DOCUMENT RESUME

ED 441 856 TM 030 891

AUTHOR Nevitt, Jonathan
TITLE Evaluating Small Sample Approaches for Model Test Statistics in Structural Equation Modeling.
PUB DATE 2000-04-24
NOTE 63p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE EDRS Price MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Error of Measurement; Goodness of Fit; Maximum Likelihood Statistics; Monte Carlo Methods; *Sample Size; Simulation; *Structural Equation Models
IDENTIFIERS Type I Errors

ABSTRACT

Structural equation modeling (SEM) attempts to remove the negative influence of measurement error and allows for investigation of relationships at the level of the underlying constructs of interest. SEM has been regarded as a "large sample" technique since its inception. Recent developments in SEM, some of which are currently available in popular SEM software programs, have lead to methods that may be more viable with small samples. While initial research investigating these methods shows promise, there is clearly the need for more thorough inspection to understand the behavior of these methods under varying experimental conditions. This study uses a Monte Carlo simulation experiment to explore the potential of these small sample methods for providing adequate model test statistics used in evaluating overall model fit. Type I error behavior and power were examined using maximum likelihood (ML), Satorra-Bentler scaled and adjusted (SB; A. Satorra and P. Bentler, 1988, 1994), residual-based (M. Browne, 1984), and asymptotically distribution free (ADF; M. Browne, 1982, 1984) test statistics. To accommodate small sample sizes, the ML and SB statistics were adjusted using a "k"-factor correction (M. Bartlett, 1950), and the residual-based and ADF statistics were corrected using modified chi-square and "F" statistics (K. Yuan and P. Bentler, 1998, 1999). Design characteristics include model type and complexity (15-variable structural equation and 21-variable confirmatory factor analysis models), sample size-to-estimated parameter ratio (n:q=1:1, 2:1, 5:1, and 10:1), and distributional form (normal and nonnormal, elliptically symmetric and asymmetric). (Contains 9 tables, 4 figures, and 33 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 441 856

Running Head: SMALL SAMPLE TEST STATISTICS IN SEM

Evaluating Small Sample Approaches for Model Test Statistics in Structural Equation Modeling

Paper presented at the Annual meeting of the American Educational Research Association.

New Orleans, LA. April 24, 2000

Jonathan Nevitt

University of Maryland

Department of Measurement, Statistics, and Evaluation

1230 Benjamin Building

College Park, MD 20742-1115

Email: jnevitt@wam.umd.edu

TM030891

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

J. Nevitt

2A

BEST COPY AVAILABLE

Abstract

Many issues in social science research may warrant investigation using latent variable methods – structural equation modeling (SEM) attempts to remove the negative influences of measurement error and allows for investigation of relationships at the level of the underlying constructs of interest. SEM has been regarded as a "large sample" technique since its inception. Unfortunately, it is often the case that large samples are either unavailable to practitioners or not practical to obtain. Thus, social scientists may find themselves turning away from latent variable modeling techniques due to apparent sample size restrictions.

Recent developments in SEM have led to methods that may be more viable with small samples, some of which are currently available in popular SEM software programs. While initial research investigating these methods shows promise, there is clearly the need for more thorough inspection to understand the behavior of these methods under varying experimental conditions. The present investigation uses a Monte Carlo simulation experiment to explore the potential of these small sample methods for providing adequate model test statistics used in evaluating overall model fit.

Type I error behavior and power was examined using maximum likelihood (ML), Satorra-Bentler scaled and adjusted (SB; Satorra &Bentler, 1988, 1994), residual-based (Browne, 1984), and asymptotically distribution free (ADF; Browne, 1982, 1984) test statistics. To accommodate small sample sizes the ML and SB statistics were adjusted using a k-factor correction (Bartlett, 1950); the residual-based and ADF statistics were corrected using modified chi-square and F statistics (Yuan & Bentler, 1998, 1999). Design characteristics include model type and complexity (15-variable structural equation, and 21-variable confirmatory factor analysis models), sample size-to-estimated parameter ratio (n:q=1:1, 2:1, 5:1, and 10:1), and distributional form (normal and, nonnormal elliptically symmetric and asymmetric).

Introduction

Structural equation modeling (SEM) has become a powerful and widely used data analytic method in psychological and behavioral research. The principal considerations in SEM are parameter estimation, evaluating model fit, and inspecting individual model parameters. Assessing global model fit via an appropriate test statistic remains paramount in the modeling process – concern over individual model parameters is pointless if an hypothesized model is not consistent with the data. This investigation focuses on the problem of testing overall model fit when sample size is small and data do not meet underlying distributional assumptions.

For a system of p measured variables, let Σ_0 represent the true population covariance matrix underlying the variables of interest. Then, a covariance structure model represents the elements of Σ_0 as functions of model parameters with null hypothesis H_0: $\Sigma_0=\Sigma(\theta)$, in which θ is a vector of q model parameters. An hypothesized model may be fit to a p x p sample covariance matrix (**S**), and for any vector of model parameter estimates ($\hat{\theta}$) the hypothesized model can be used to evaluate the model-implied covariance matrix, $\Sigma(\hat{\theta})=\hat{\Sigma}$. The goal in parameter estimation is to obtain a vector of parameter estimates such that $\hat{\Sigma}$ is as close to **S** as possible (i.e., the model is consistent with the observed data). The disparity between $\hat{\Sigma}$ and **S** is measured by a discrepancy function which is sought to be minimized – the maximum likelihood (ML) function is the most commonly employed discrepancy function for estimating model parameters and is defined as: $\hat{F}_{ML} = \ln|\hat{\Sigma}|-\ln|\mathbf{S}|+tr(\mathbf{S}\hat{\Sigma}^{-1})-p$ (see, e.g., Hayduk, 1987), where tr(.) is an operator that sums the elements on the main diagonal of a matrix (i.e., obtains the trace of the matrix.).

ML estimation is ubiquitous in SEM and has become the default estimation method in popular SEM software programs such as AMOS (Arbuckle, 1997), EQS (Bentler, 1996), and LISREL (Jöreskog

& Sörbom, 1996). The popularity of ML estimation stems from its desirable statistical properties – ML yields unbiased, consistent, and efficient parameter estimates, and provides a model test statistic [$T_{ML}=(n-1)\hat{F}_{ML}$, for a sample of size n] for assessing the adequacy of an hypothesized model. These statistical properties inherent in ML are asymptotic results derived from the theoretical behavior of ML under the assumption of multivariate normality and infinitely increasing sample size. Thus for a system of p measured variables (yielding $p^*=p(p+1)/2$ unique variances and covariances) and q model parameters to be estimated, asymptotic normal theory establishes that T_{ML} follows a central chi-square distribution with p^*-q degrees of freedom (df).

Practitioners using SEM are advised to obtain large sample sizes to ensure that the asymptotic properties underlying ML estimation are reasonably well met. However, real-world data often yield small sample sizes, potentially jeopardizing the approximation to referenced distributions. (A distinct but related issue is that even when sample sizes are large, nonnormal data forms will distort the T_{ML} test statistic. As discussed later, sample size and distributional form have been linked together with respect to developing accurate test statistics for assessing global model fit in SEM.) Still, large samples are not required per se to obtain stable parameter estimates using ML; it has been suggested that a potential lower bound might be as small as n only slightly greater than p (Bentler & Yuan, 1999). While model parameters can be estimated with small samples, model testing may be compromised. Unfortunately, for assessing global model fit, no statistical theory exists to determine how large sample sizes must be before a reasonable approximation to referenced distributions is attained. Key studies have examined the propensities of T_{ML} under small sample conditions, as well as the behavior of the test statistic with nonnormal data forms and are briefly reviewed below. Then, test statistics for evaluating model fit that may be appropriate under small sample conditions are considered.

Maximum Likelihood Estimation – Impact of Small Sample Sizes and Nonnormal Data

Empirical research in SEM, with respect to issues of sample size and normality, has generated a large body of literature and an understanding of the behavior of ML estimation (and its associated test statistic) with realistic data forms (Bentler & Yuan, 1999; Boomsma, 1982, 1985; Curran, West, & Finch, 1996; Fouladi, 1998, 1999; Gerbing & Anderson, 1985; Hu & Bentler, 1995; Hu, Bentler, & Kano, 1992; Nevitt & Hancock, in press-b; Tanaka, 1984, 1987; Yuan & Bentler, 1997, 1998, 1999). Early work investigating sample size requirements in SEM with normal data was conducted by Boomsma (1982, 1985), and by Gerbing and Anderson (1985), with findings that at small sample sizes ML estimation may lead to problems with model non-convergence and improper solutions, and that for convergent models Type I error rates become inflated. Tanaka (1984) investigated the effects of sample size in latent variable models, varying sample size as well as distributional form of the data and found marginal degradation in parameter estimates and standard errors, and inflation in model fit statistics at samples of size n=100. Tanaka concluded that the type of sample size concerns noted in the Boomsma (1982, 1985) and Gerbing and Anderson (1985) studies become more acute under conditions of nonnormality. More recent empirical evidence supports the distortion of T_{ML} at small samples and nonnormal data conditions (Bentler & Yuan, 1999; Curran et al., 1996; Fouladi, 1998, 1999; Hu et al., 1992 Nevitt & Hancock, in press-b).

Tanaka (1987) briefly summarizes the extant research at the time with respect to the sample size issue in SEM. One emerging theme is that sample size requirements in SEM should not be judged in an absolute sense, but with respect to the ratio of observations in the sample to the number of parameters estimated in the model (n:q). The SEM community has subsequently adopted (although rather loosely) an n:q=5:1 ratio for ML estimation (Bentler, 1996, p. 6; Bentler & Chou, 1987). The findings reported in Boomsma (1982) and Tanaka (1987) suggest that for T_{ML} to be effective in evaluating model fit even larger ratios may be necessary.

To summarize, empirical results indicate that small samples (and nonnormal data) may lead to problems with respect to model convergence and improper solutions, but generally yield unbiased parameter estimates for models that converge. A more concerning issue is evaluating model fit using T_{ML}; findings indicate that at small samples T_{ML} becomes inflated when fitting properly specified models (i.e., Type I error rates rise above the nominal level). Moreover, violation of the normality assumption underlying ML estimation exacerbates this tendency for T_{ML} to inflate. To combat the distortion in the model fit statistic at small samples and nonnormal data two fundamentally divergent approaches have emerged: abandoning ML for distribution-free estimation methods, and applying ML for parameter estimation but using an adjusted test statistic to account for sample size and nonnormality. These two approaches are considered next.

Alternative Estimation Methods and Associated Test Statistics

Browne (1982, 1984) developed a procedure known as asymptotically distribution-free estimation (ADF) which is not tied to a particular distributional form of the observed data. Browne proposed estimating model parameters by minimizing $\hat{F}_{ADF}$, a general function for expressing the weighted differences between observed and model-implied (co)variances. The ADF estimator yields asymptotically unbiased and consistent estimates for model parameters without the need for distributional assumptions on the data; it also yields a test statistic for assessing overall model fit. Under the null hypothesis, the associated test statistic $T_{ADF} = (n\text{-}1)\hat{F}_{ADF}$ asymptotically follows a central chi-square distribution with $p^*\text{-}q$ df.

At large sample sizes ($n \geq 5{,}000$) T_{ADF} operates as expected, yielding observed Type I error rates at the nominal level (Chou, Bentler, & Satorra, 1991; Curran et al., 1996; Hu et al, 1992). However, with large models and small to moderate sample sizes ADF is problematic; it tends to yield high rates of non-

convergence and improper solutions when minimizing $\hat{F}_{ADF}$ (a problem much less frequently encountered when minimizing $\hat{F}_{ML}$). Moreover, T_{ADF} rejects true models too often, yielding Type I error rates as high as 68% (Curran et al., 1996) under some conditions. Such problems with ADF estimation at small samples have been linked to instability in the estimated fourth-order-moment weight matrix requisite in the method.

ADF's ability to yield a correct test statistic under nonnormal conditions at large sample sizes inspired Yuan and Bentler (1997, 1999) to develop corrections to T_{ADF} for small sample sizes. They note that in the regression literature (a data-analytic paradigm in which small samples are less problematic as compared to SEM) cross products of model residuals are often used for estimating asymptotic covariances. With such motivation in place they proposed an improved estimator for the weight matrix in $\hat{F}_{ADF}$ and note that incorporating this estimator into ADF is tantamount to applying a multiplicative scaling constant to the original ADF test statistic yielding $T_{ADF1}=\{1/[1+(T_{ADF}/(n-1))]\}T_{ADF}$. T_{ADF1} follows a central chi-square distribution with p^*-q df and is asymptotically equivalent to T_{ADF}. Yuan and Bentler (1997) conjecture that this adjusted test statistic may yield improved Type I error performance at reduced sample sizes and demonstrate such performance empirically (Yuan & Bentler, 1999).

Further motivated to improve small sample performance associated with T_{ADF}, Yuan and Bentler (1999) proposed another modification to the ADF test statistic. They appeal to the F distribution rather than the chi-square distribution for this modified test statistic, noting that test statistics are often referred to the former distribution rather than the latter when sample sizes are not large. This motivation inspired them to investigate a transformation of T_{ADF} similar to the transformation applied to Hotelling's T^2 statistic in MANOVA. They proposed rescaling T_{ADF} to $T_{ADF2}=[n-(p^*-q)]/[(n-1)(p^*-q)]T_{ADF}$, referring the test statistic to an F distribution for hypothesis testing with degrees of freedom (p^*-q) and $[n-(p^*-q)]$ for numerator and denominator, respectively. As with T_{ADF1}, Yuan and Bentler (1999) hypothesize and empirically demonstrate that the T_{ADF2} adjusted test statistic may yield improved Type I error

performance at small sample sizes. Inspired by the limited yet promising simulation results of Yuan and Bentler (1997, 1999) the adjusted ADF test statistics have been made available to practitioners: T_{ADF1} is currently available in EQS 5.7 (Bentler, 1996), and T_{ADF2} will be incorporated into the future release of EQS 6.0.

Maximum Likelihood-Based Test Statistics: Corrections to T_{ML}

The ADF test statistic and its corrected forms, T_{ADF1} and T_{ADF2}, control for nonnormal sample data by employing distribution-free estimation. Another approach to controlling for nonnormal data (and potentially small samples) is to estimate model parameters using ML and then assess model fit via a test statistic that has been corrected for nonnormality and sample size. These corrections take two basic forms: adjusting T_{ML} for sample size and nonnormality, and constructing distribution-free test statistics based on ML-estimated model parameters. Both approaches are considered here.

The most well known and widely investigated corrections to T_{ML} have been developed to accommodate the inflation seen in T_{ML} under nonnormal data conditions. Satorra and Bentler (1988, 1994) pioneered this approach with two modifications to T_{ML} that make adjustments based on the model and the degree of nonnormality in the sample data (generically referred to as SB test statistics in this report). One correction involves a multiplicative scaling of T_{ML} (T_{SB1}), while the other is an adjusted test statistic (T_{SB2}) incorporating both a scaling of T_{ML} and a degree of freedom adjustment. Simulation research has indicated that both SB adjustments to T_{ML} can be effective in controlling Type I error rates under some experimental conditions (Chou et al., 1991; Curran et al., 1996; Fouladi, 1998, 1999; Hu et al., 1992; Nevitt & Hancock, in press-b; Yuan & Bentler, 1998). The scaled T_{SB1} is currently available in EQS 5.7 (Bentler, 1996) and LISREL 8.3 (Jöreskog & Sörbom, 1996).

Fouladi (1998, 1999) proposed another correction to T_{ML}, combining a small-sample correction developed by Bartlett (1950) with the SB corrections for nonnormality. Bartlett (1950) offered a k-factor

BEST COPY AVAILABLE

small sample correction to test statistics within the context of explora
(1998) proposed applying the correction to T_{ML} in SEM to improve i
normal data conditions and Fouladi (1999) proposed applying it to T
multiplicative scaling constant applied to the test statistics is of the fo
k is the number of latent factors in the model of interest. The correct
square distribution with p^*-q df. Fouladi (1999) investigated the sma
corrected forms of T_{ML}, T_{SB1}, and T_{SB2} using a 6-variable path model
Reported findings in those studies showed this scaling correction to b
error rates under some experimental conditions. In particular, the k-f
the strongest performance under the combination of small sample siz

Residual-Based Test Statistics

When real data give rise to nonnormal small samples, neither
test statistics nor adjustments to T_{ML} present ideal solutions. For sma
frequently leads to model convergence failures and improper solution
A more imposing limitation to ADF estimation is the minimum samp
lower bound for sample size is $n=p^*$ (Bentler & Yuan, 1999). Thus,
indicator variables, minimum sample size requirements for ADF esti

Corrections to T_{ML} with nonnormal small samples also have t
correction as proposed by Fouladi (1998, 1999) is mostly uninvestiga
with no latent variables ($k=0$) are encouraging, but no studies have be
nature of k-factor corrected test statistics with latent variable models
correction is based on a priori knowledge of the true number of facto
that may be hypothesized but not necessarily known. The SB adjustr

of exploratory factor analysis (EFA). Fouladi
improve its small sample performance under
ing it to T_{SB1} and T_{SB2} for nonnormal data. The
is of the form c=1-(2p+4k+5)/6(n-1), in which
he corrected test statistic is referred to a chi-
ed the small sample behavior of k-factor
ath model with no latent variables (i.e., k=0).
ection to be effective in controlling Type I
lar, the k-factor-corrected form of T_{SB2} showed
sample size and nonnormal data.

es, neither ADF estimation and its (corrected)
s. For small samples, ADF estimation
per solutions, especially when models are large.
mum sample size requirement; the absolute
99). Thus, for models with large numbers of
r ADF estimation may be quite large.
also have their limitations. The k-factor
uninvestigated; results based on a path model
dies have been conducted characterizing the
ple models (i.e., k>0). Furthermore the
per of factors underlying the data, information
SB adjustments to T_{ML} have shown adequate

performance in practice but have been criticized for lacking known asymptotic results with nonnormal data. Thus, T_{SB1} and T_{SB2} may behave unpredictably under varying distributional forms of sample data.

Facing such issues Yuan and Bentler (1998), and Bentler and Yuan (1999) turned to a mostly ignored residual-based test statistic (T_B) originally developed by Browne (1984). They note that T_B may be applied to any consistent estimator for θ, even when data are not normally distributed, and argue that ML is a prudent choice (because ML parameter estimates tend to be unbiased). They also note that T_B is asymptotically distributed as a central chi-square with p^*-q df. The test statistic is currently available to practitioners within LISREL 8.3 (Jöreskog & Sörbom, 1996).

The advantage to T_B is that it is a distribution-free test statistic with theoretical elegance (i.e., it follows a known sampling distribution without assuming multivariate normality of the data) not found with T_{SB1}, T_{SB2}, or the k-factor corrected test statistics. Additionally, Bentler and Yuan (1999) indicate that T_B also has smaller minimum sample size requirements than T_{ADF} and its adjusted forms if ML estimation is used; the minimum sample size for T_B is $n=p^*-q+1$ (i.e., the model df + 1). Thus, T_B may be obtained for sample sizes smaller than those required for ADF estimation. Yuan and Bentler (1998), and Bentler and Yuan (1999) showed that T_B has a propensity to over-reject true models at sample sizes smaller than n=5000 – Type I error performance remarkably similar to that of the unadjusted T_{ADF} test statistic.

The poor performance of T_B at small samples inspired Yuan and Bentler (1998) to propose corrections to the test statistic analogous to the small sample corrections applied to the ADF test statistic. First, they proposed applying a multiplicative scaling constant to T_B yielding $T_{B1}=\{1/[1+(nT_B)/(n-1)^2]\}T_B$, with T_{B1} following a central chi-square distribution with p^*-q df. Yuan and Bentler (1998) also note that the equation for T_B takes on a quadratic form also reminiscent of Hotelling's T^2 statistic. They proposed rescaling T_B to $T_{B2}=\{[n-(p^*-q)]/[(n-1)(p^*-q)]\}T_B$, referring the test statistic to an F distribution for hypothesis testing with degrees of freedom (p^*-q) and $[n-(p^*-q)]$ for numerator and denominator,

respectively. Yuan and Bentler (1998) note that T_{B1} and T_{B2} are asymptotically equivalent to T_B and thus are asymptotically distribution free. However, at small to moderate sample sizes, they expect an improvement in model rejection rates for the scaled forms of the test statistic; such expectations were demonstrated empirically by Yuan and Bentler (1998) and Bentler and Yuan (1999).

Rationale for Further Investigation

While the test statistics discussed above yield promising results, there is clearly the need for further research to more fully characterize the nature of small sample test statistics in SEM. First, no studies have investigated the $\underline{k}$-factor correction with latent variable systems (for which $\underline{k}>0$), and thus it is unknown if the improvement afforded to T_{ML} for normal data and the improvements to T_{SB1} and T_{SB2} with nonnormal data conditions using the $\underline{k}$-factor multiplier will hold under such models. If an improvement with latent variable models is realized using this correction, then determining a lower bound on sample size requirements for the effective use of the $\underline{k}$-factor correction could facilitate making recommendations to practitioners.

In addition to furthering our knowledge of the potential of the $\underline{k}$-factor correction, much remains to be learned regarding the behavior of the ADF and residual-based adjusted test statistics, T_{ADF1}, T_{ADF2} and T_{B1}, T_{B2}. Bentler and Yuan (1999) note that much of the characterization of these test statistics focuses only on a specialized model under a very narrow range of conditions. They call for continued research, investigating these test statistics under a greater variety of conditions, including variations in the number of variables in the system and the type of model examined. Furthermore, for sample sizes in which ADF estimation is possible, no research exists that compares the relative performance of T_{ADF1} and T_{ADF2} against their respective residual-based analogs, T_{B1} and T_{B2}. Thus for moderate sample sizes, no clear choice exists between the ADF-based test statistics and the residual-based test statistics (using ML estimation). Finally, Bentler and Yuan (1999) comment on the power associated with the residual-based

test statistics to reject improperly specified models. They did not investigate the power of T_{B1} and T_{B2} in their study, but they note that characterizing the power of these test statistics is important in establishing their behavioral profile and requires further investigation. This current investigation seeks to address these important issues, filling the void in our understanding of small sample test statistics in SEM.

Method

The present study examines the small sample performance of model test statistics used for evaluating overall model fit in SEM. A Monte Carlo simulation experiment was conducted manipulating design aspects involving model type, distributional form, and sample size. These aspects of the design as well as issues addressing the test statistics that were examined, the summary measures employed, the number of replications performed, the robustness criteria for evaluating test statistics, and follow-up power investigations are fully elaborated below.

Test Statistics Examined

The present study examines test statistics for evaluating global fit in SEM that, based on previous investigations, have shown promise for practitioner use under realistic data conditions. With respect to corrected forms of T_{ML}, the SB corrections have performed well at moderate sample sizes with nonnormal data and at small sample sizes when combined with the k-factor correction, and thus deserve further examination. For normal data conditions, the k-factor correction to T_{ML} demonstrated potential robustness at very small sample sizes. In light of these findings, the following T_{ML} corrected statistics were investigated in the present study: T_{SB1}, T_{SB2}, $T_{ML\text{-}k}$, $T_{SB1\text{-}k}$, and $T_{SB2\text{-}k}$ (test statistics labeled with a trailing k indicate the test statistic has been subjected to the k-factor correction). The uncorrected T_{ML} test statistic was analyzed under all conditions in this study, providing a benchmark for comparison against its adjusted forms. The ADF adjusted test statistics T_{ADF1} and T_{ADF2} also have yielded promising results for

nonnormal small sample data, as have the residual-based adjusted test statistics T_{b1} and T_{B2}. These test statistics certainly are worthy of continued investigation. As with the corrected T_{ML} test statistics, the unadjusted T_{ADF} and T_B test statistics were also analyzed in the present study for benchmark comparison purposes.

Population Models and Model Specifications for Type I Error Analyses

Two population models were established to investigate the Type I error behavior and power of the test statistics under investigation. The layout of the population models and their correct model specifications used for fitting sample data (for Type I error analysis) are described here; misspecified models for power analyses are described later in a separate section. For each population model, two model specifications were examined creating four model specification conditions (all properly specified) that systematically vary model size (p), model complexity (q) and model df. These design characteristics were manipulated to address the issue raised by Yuan and Bentler (1998) of test statistic failure and model complexity. For each population model, standard path tracing techniques were used to obtain the associated true population covariance matrix $\Sigma_0=\Sigma(\theta)$ from which simulated raw data were drawn.

Population Model A is shown in Figure 1. It is a 5-factor 15-variable structural equation (SE) model with three indicator variables loading on each factor. The structural portion of the model contains two exogenous and three endogenous factors, and is structured to have population values as indicated in the figure.

Insert Figure 1 about here

Model specification A1 fits sample data generated via Population Model A to an unconstrained correctly specified model requiring a total of q=35 parameter estimates yielding 85 df. Figure 1 shows the path diagram representing the model to which sample data were fit to yield Model specification A1.

Asterisks on the diagram indicate parameters requiring estimation; values in the diagram represent fixed parameters.

Model specification A2 (Figure 1) also fits sample data generated via Population Model A to a correctly specified model, but this model is constrained in the measurement portion of the model requiring a total of q=18 parameters to be estimated and yielding 102 df. For Model A2, all free loadings within a factor were constrained to equality, as were the three error variances associated with the factor. (Because the true population model contains loadings and error variances within each factor that are equal, constraining them to equality when fitting sample data does not introduce any lack of fit into the model and as such is correctly specified.)

Population Model B was established to investigate a larger latent variable system. The model is displayed in Figure 2 (along with Model specifications B1and B2), and is a 7-factor 21-variable CFA model. Population Model B includes non-zero error covariances among the three groups of error terms E1, E4, E7, E10, E13, E16, E19; E2, E5, E8, E11, E14, E17, E20; and E3, E6, E9, E12, E15, E18, E21; three additional non-zero error covariances were also established for E1,E2; E10,E11; and E19,E20 for a total of 66 non-zero error covariances.

Insert Figure 2 about here

Model specification B1 fits sample data generated via Population Model B to a correctly specified yet constrained model requiring a total of q=35 parameters to be estimated and yielding 196 df. Model specification B2 also fits sample data generated via Population Model B to a correctly specified model, but this model was established such that many model parameters require estimation yielding relatively small model df (as compared to Model B1). Model B2 releases the equality constraints on the measurement portion of the model that were imposed in Model B1 and estimates all of the non-zero error covariances in Population Model B for a total of q=129 and 102 df. Table 1 summarizes for each model

specification, the number of variables (p) and non-redundant variances and covariances (p*) in the system, the number of parameters requiring estimation (q), and the model df.

Insert Table 1 about here

Sample Size Conditions

With respect to minimum sample size requirements, previous studies have pointed out that the most effective barometer by which to gauge sample size adequacy may be the n:q ratio as opposed to an absolute sample size criterion (e.g., Chou & Bentler, 1987; Tanaka,1987). Clearly the models under examination in the present investigation differ with respect to model size (p) and model complexity (q), suggesting that holding sample sizes (in an absolute sense) constant across models is not useful. Rather, sample size-to-parameter ratio conditions were established and held constant across the four models, with each model then taking on different absolute sample sizes that reflect the n:q ratio conditions. For all four models ratios of n:q=1:1, 2:1, 5:1, and 10:1 were examined. In addition each model was also inspected under a sample size condition at which the residual-based and ADF test statistics are at their respective minima (n=df+1for the residual based fit measures and n=p* for the ADF statistics). Thus each model was examined under a total of six sample size conditions.

Test statistics based on T_{ML} are not limited to any sample size constraints per se and were examined across all six of the sample size conditions. The residual-based test statistics were only examined under sample size conditions in which n≥df+1 while the test statistics based on ADF estimation were only examined under sample size conditions in which n≥p*. Because absolute sample sizes varied across models and the residual-based and ADF test statistics have minimum sample size constraints in an absolute sense, the number of sample size conditions under which these test statistics were examined varied across models. Table 1 charts each combination of model and sample size condition and the associated absolute sample size for that cell in the design of the experiment.

Data Generation and Distributional Forms

Three multivariate distributions were examined, each established through the manipulation of univariate skew and kurtosis of the measured variables. All manifest variables were drawn from the same univariate distribution for each data condition. Distribution 1 is multivariate normal with univariate skew and kurtosis both set equal to 0. (Note that normality is defined here, as is commonly done in practice, by using a shifted kurtosis value of 0 rather than a value of 3.) Distribution 2 represents an elliptical distribution – data are nonnormal but symmetric with univariate skew of 0 and kurtosis of 6.0. Distribution 3 is nonnormal and asymmetric with univariate skew of 3.0 and kurtosis 21.0.

Simulated data matrices were generated in GAUSS (Aptech Systems, 1996) to achieve the desired levels of univariate skew, kurtosis, and covariance structure underlying a given population model. Multivariate normal and nonnormal data were generated via the algorithm developed by Vale and Maurelli (1983), which is a multivariate extension of the method for simulating nonnormal univariate data proposed by Fleishman (1978). The programming to generate the simulated data for the present investigation has been tested and verified for accuracy; the programs were scrutinized externally and are now available to the general public (Nevitt & Hancock, 1999). Seeding for the simulations was done carefully to ensure that no two seed values for Models A1 and A2, or for Models B1 and B2 were identical, thus ensuring that unique data sets were always drawn from a particular population model.

Computation and Verification of Test Statistics

All modeling of simulated data was performed using EQS 5.7b (Bentler, 1996). The T_{ML} and T_{ADF} test statistics were captured directly from EQS, as were the Jacobian matrix of partial derivatives and the residual matrix (of discrepancies between observed and model-reproduced variances/covariances). All other test statistics were calculated using GAUSS. Because most of the test statistics in the present investigation were constructed outside of EQS, careful steps were taken to verify

the accuracy of the fit statistics whenever possible. Calculation of the components of the SB test statistics were verified using EQS 5.7b (Bentler, 1996), and the residual-based test statistic T_B was verified using LISREL 8.3 (Jöreskog & Sörbom, 1996). The ADF-corrected test statistics were also verified for accuracy; T_{ADF1} values were compared against output from EQS 5.7b (Bentler, 1996) and T_{ADF2} statistics were matched up against output from a pre-release of EQS 6.0.

Design

For the investigation of Type I error rates, the present study fully crossed the between-cells aspects of the design to yield 72 cells (i.e., four model specifications by six sample size conditions by three distributional forms). For each cell in the design, independent data sets were drawn from the associated population model, distributional form, and sample size, and fit to the appropriate model specification using EQS. The same simulated data sets were not used to fit different model specifications (i.e., across Models A1 and A2 or, B1 and B2) even though they might have been drawn from the same population model, distributional form, and sample size. Such an approach was adopted to prevent any potential dependencies among the results across Models A1 and A2 or, across Models B1 and B2. The number of iterations to convergence was set at 200 for ML estimation; EQS 5.7b (Bentler, 1996) automatically limits the maximum number of iterations for ADF estimation to 30. Start values for fitting simulated samples of data to models were assigned using the true population parameter values.

Simulated data were subjected to parameter estimation using ML and ADF to obtain the T_{ML}-based, residual-based and ADF-based test statistics. For each data set all test statistics were computed, except for those data conditions in which the absolute n was less than the minimum sample size for the residual-based or ADF test statistics. For each cell in the design the target number of replications was set at 2000 for the Type I error analysis. The target number of 2000 replications refers to the number of test statistics analyzed per cell of the design (2000 for each of the twelve test statistics investigated) rather

than referring to the number of data sets subjected to modeling. Once a test statistic reached its target number of replications a counter shut off that test statistic from subsequent data modelings. For some cells in the design (particularly cells at the ADF minimum sample size) large numbers of data sets were needed to obtain the target number of test statistic replicates. In the worst case scenario (Model A1 with asymmetric nonnormal data at $\underline{n}=\underline{p}^*$) 21,704 data sets were required to yield the target number of replications for that cell because of problems with ADF non-convergence and improper solutions.

Summarizing Results and Defining Test Statistic Robustness

Study results for the properly specified models were analyzed in terms of empirical Type I error rates. Yuan and Bentler (1998, 1999) note that although adherence to the referenced sampling distribution is important, the primary concern for hypothesis testing is with respect to the tail-behavior of the test statistic (i.e., in terms of model rejection rates). As such, the proportion of model rejections across each cell's replications are reported (in percentage form to facilitate readability) for each study condition and are used to gauge test statistic robustness; all hypothesis testing was performed at the 5% level of significance.

Robustness with respect to Type I error rates is evaluated using both an expected 95% confidence interval (CI) about $\underline{\alpha}=.05$ and Bradley's liberal criterion (Bradley, 1978). The expected 95% CI is useful in determining if a given test statistic's Type I error rate appears to deviate from the nominal level. Given the 2000 replications per condition in this investigation and $\underline{\alpha}=.05$ the 95% CI (in percentage form) evaluates to [4.04%, 5.96%] by applying the formula: $[.05 \pm 1.96(.05 \times .95/2000)^{1/2}] \times 100\%$. A test statistic yielding an empirical rejection percentage lying outside this interval is considered as having a true Type I error rate other than 5%.

Test statistics are frequently considered useful (robust) if their Type I error propensity is at least reasonably maintained, albeit not necessarily at the nominal level of significance. Bradley (1978) sought

to control interpretation of test statistic robustness by proposing a quantification of the property. Using Bradley's liberal criterion, an estimator is considered robust if it yields an empirical model rejection rate within the interval [.5α, 1.5α]. Using α=.05, this interval for robustness of rejection rates (in percentage form) is [2.5%, 7.5%]. While this interval is certainly wider than the 95% CI given above, the two intervals represent fundamentally different aspects of test statistic performance. The 95% CI assesses deviation from exactly the nominal level; the robustness interval defines acceptable departure from the nominal level.

Follow-up Power Analyses

While the Type I error propensity of the test statistics investigated is the central focus in the present study, the power of a test statistic to reject an improperly specified model is also considered. This investigation takes a two-step approach to evaluating test statistic utility. In the first step, the Type I error propensity is evaluated. Using Bradley's liberal criterion, test statistics under study conditions that yielded empirical Type I error rates above the upper bound of 7.5% were culled from subsequent power analyses under those same conditions. Such an approach is reasonable because if Type I error rates are inflated then the power of that test statistic to reject models is artificially inflated and thus not informative. Test statistics with model rejection rates within the robustness interval, or below the lower bound of 2.5%, were examined in a follow-up power analysis. Test statistics that yielded empirical Type I error rates below the lower bound of robustness were retained for power analysis, rather than being culled, because these test statistics may potentially yield acceptable power with respect to rejecting misspecified models.

For each test statistic that yielded an observed Type I error rate that was within (or below) the robustness interval, a new series of 1000 replicates were generated using the same population model for that cell in the study, and at the same distributional form and sample size condition. These new data sets were then fit to an improperly specified model. A target of 1000 replicates per test statistic was deemed

sufficient for the power analyses because the usefulness of the power estimates may not hinge upon as high a degree of precision as was deemed necessary for the Type I error analysis.

For each of the four model specifications investigated in the Type I error analysis, an associated misspecified model was established for purposes of follow-up power analyses. Each misspecified model was constructed to yield model df closely matching the model df under its companion properly specified model. Because the misspecified models did not perfectly match in model df against their companion properly specified models it was not possible to maintain the exact same absolute sample sizes and n:q ratios in the cells of the power analysis as compared to the Type I error analysis. Thus the absolute sample sizes examined in the Type I error analysis were preserved in the power analysis (rather than the n:q ratios), an approach consistent with other research studies (cf. Yuan & Bentler, 1998, 1999).

Figure 3 shows the improper model specifications corresponding to Models A1 and A2 under the power analysis. Sample data drawn from the 5-factor Population Model A were fit the to the 3-factor misspecified SE models as shown. In misspecified Model A1 all loadings and error variances are freely estimated (except for the loading of V15 on F3 that is fixed to achieve identification); for misspecified Model A2 the measurement model is constrained, estimating only one factor loading and error variance per factor. Model df for these misspecified models are 87 and 110, respectively, which are reasonably close to the 85 and 102 df for their respective properly specified models.

Insert Figure 3 about here

Misspecified Models B1 and B2 are given in Figure 4; sample data drawn from the true 7-factor CFA Population Model B were fit to the misspecified 4-factor CFA model. Misspecified Model B1 has a constrained measurement model, estimating only 1 factor loading and error variance per factor to yield 217 df; misspecified Model B2 estimates all loading parameters and error variances, as well as the 66

error covariances estimated in the properly specified Model B2, yielding 117 df. Note that model df for these misspecified models diverge slightly more from the df in their companion properly specified models (compare against the df of 196 and 102 for properly specified Models B1 and B2, respectively).

Insert Figure 4 about here

The proportion of model rejections (again tabled as percentages) across the replicates represents an empirical estimate of the power associated with a given test statistic. The rationale for such a power analysis is that for a set of test statistics which demonstrate Type I error robustness under a particular study condition, a clear choice (and hence recommendations to practitioners) for the best statistic may not be obvious. With the additional information regarding the power of the statistic in hand, more definitive recommendations to practitioners can be made.

Results

Non-convergence, Improper Solutions, and Test Statistic Failure Rates

Percentages of non-convergence and improper solutions for the ML and ADF estimation methods, as well as failure rates for the SB and residual-based test statistics were carefully scrutinized under all models investigated. Improper solution and test statistic failure rates were based only on those data sets that yielded converged solutions. Overall, non-convergence and improper solutions for ML and ADF were a function of distributional form and sample size, with increasing departure from multivariate normality and decreasing sample size leading to systematically increasing rates of non-convergence and improper solutions. For every simulated data set in this investigation the SB test statistics never failed (i.e., the triple-product matrix requisite for the computation of the SB-based test statistics was successfully inverted every time) leading to a constant 0% failure rate for these test statistics. Like the SB

test statistics, the T_B test statistic also had 0% failure rates for inverting its necessary product matrix, but was flagged as failing (and discarded) when a computed value of T_B exceeded a value of T_B=900,000. Descriptive results for non-convergence and improper solutions are given below; summary tables are available from the author upon request.

Non-convergence rates for ML under Model A1 were minimal across distributional forms and sample sizes, never exceeding 2.5%. Rates of improper solutions for ML at sample size conditions of $\underline{n}:\underline{q} \geq 2:1$ were usually 2.5% or less, with a notable exception under asymmetric nonnormal data where improper solution rates climbed to 10-16.5%. At the $\underline{n}:\underline{q}=1:1$ sample size conditions, rates of improper solutions were considerably higher at about 30% for normal and elliptically symmetric nonnormal data, and about 62% with asymmetric nonnormal data. The residual-based test statistic, T_B, had failure rates only at its minimum sample size condition of $\underline{n}=\underline{df}+1$ with failure percentages of about 7%; distributional form did not impact the rate of test statistic failure for T_B under Model A1. Convergence results for ADF were especially problematic at ADF's smallest sample size of $\underline{n}=\underline{p}^*$ with non-convergence reported at 25-40% across the three distributional forms. Instances of improper solutions were even more frequent at $\underline{n}=\underline{p}^*$ with as many as 85% of all converged data sets yielding improper solutions with asymmetric nonnormal data.

For the constrained Model A2 non-convergence and improper solution rates were similar in pattern to those found for Model A1 but were also comparatively attenuated. For ML estimation, models rarely failed to converge and improper solutions were seen primarily only at the $\underline{n}:\underline{q}=1:1$ sample size condition, with a worst case of about 27% under asymmetric nonnormal data. The T_B test statistic was somewhat worse behaved (as compared to its failure rates under Model A1) delivering approximately 10% failure rates, again only at the $\underline{n}=\underline{df}+1$ sample size. ADF estimation was again problematic at the $\underline{n}=\underline{p}^*$ sample size with 20-30% non-convergence rates and 38-65% rates of improper solutions.

With the constrained CFA Model B1, ML estimation always yielded converged model fittings that only occasionally delivered improper solutions. In contrast to failure rates for T_B under Models A1 and A2, Model B1 was problematic. Failure rates for T_B at $\underline{n}=\underline{df}+1$ were substantially higher and increased monotonically with increasing departure from multivariate normality. With asymmetric nonnormal data the failure rate for T_B was nearly 50% at the $\underline{n}=\underline{df}+1$ sample size. ADF estimation was not particularly problematic for Model B1 at its smallest sample size of $\underline{n}=p^*$ with respect to model convergence. However, at $\underline{n}=p^*$ ADF again yielded large rates of improper solutions with rates of about 45% and 87% under elliptical and asymmetric nonnormal data, respectively.

ML estimation for the unconstrained Model B2 again provided perfect convergence at all sample sizes and distributional forms. Rates of improper solutions using ML under this model were elevated (as compared to Model B1), especially with nonnormal data at the smallest sample size ($\underline{n}=103$) where improper solution rates ranged from 20-45%. The T_B test statistic behaved as it had with Models A1 and A2; T_B failed only at the smallest sample size of $\underline{n}=\underline{df}+1$ with stable failure rates (across distributional forms) just below 9%. ADF estimation was again problematic for Model B2 at the smallest sample size condition of $\underline{n}=p^*$, especially with respect to rates of improper solutions that ranged from about 28% to 44% under nonnormal data conditions.

Type I Error Rates

For the properly specified Model A1, Table 2 presents empirical Type I error rates for the test statistics investigated. Tables 3-5 present parallel results for Models A2, B1, and B2, respectively.

Insert Table 2 about here

Model rejection rates in Table 2 for T_{ML} are inflated at the small sample size conditions and monotonically decrease with increasing sample size. Notice that with normal data T_{ML} is robust only at

the n:q=10:1 sample size condition, while the Bartlett k-factor correction to T_{ML} shows dramatic improvement to Type I error control with model rejection rates within the robustness interval at sample size conditions of n:q≥5:1; at smaller sample size conditions $T_{ML\text{-}k}$ delivered attenuated model rejection rates with normal data. While alternatives to T_{ML} are not usually employed with normal data it is interesting to note the strong performance of the k-factor-corrected SB scaled test statistic, $T_{SB1\text{-}k}$, and the SB adjusted test statistic, T_{SB2}. Both test statistics yielded Type I error rates that are within the robustness interval even at the smaller sample size conditions.

Moving from normal data to the elliptically symmetric nonnormal data (Distribution 2) one finds in Table 2, as expected, inflated Type I error rates associated with T_{ML}. Applying the k-factor correction to T_{ML} here alleviates the inflation in the test statistic over most of the sample size conditions with the exception of the n:q=1:1 sample size condition where $T_{ML\text{-}k}$ over-corrects. The T_{SB1} test statistic controls Type I error rates only at the largest sample size condition of n:q=10:1. Applying the k-factor correction to T_{SB1} yields a test statistic that conforms in model rejection rate at all sample sizes investigated. In fact, at all but the smallest sample size condition, the empirical Type I error rates for $T_{SB1\text{-}k}$ are within the 95% CI for the 5% level of significance. The T_{SB2} test statistic yields model rejection rates in Table 2 with elliptical data that are generally below the 2.5% lower bound of the robustness criterion. Applying the k-factor correction to T_{SB2} tends to over-correct even more so, yielding zero or near zero Type I error rates over the range of sample sizes investigated. The T_B and T_{ADF} test statistics tend to mirror one another in Model A1 with elliptical data. T_{B1} and T_{ADF1} also closely follow one another yielding suppressed Type I error rates at sample size conditions below n:q=10:1. The performance of the F statistics T_{B2} and T_{ADF2} do however diverge with elliptically nonnormal data. T_{B2} yielded Type I error rates within the robustness criterion at sample sizes greater than its minimum n=df+1. T_{ADF2} on the other hand performs correctly only at the largest sample size condition of n:q=10:1, with attenuated empirical Type I error rates associated with the test statistic.

BEST COPY AVAILABLE

Turning now to the results for the asymmetric nonnormal data (Distribution 3) given in Table 2, findings indicate that this more extreme form of nonnormality has an even greater impact on the T_{ML}, $T_{ML\text{-}k}$, and T_{SB1} test statistics as compared to the elliptically symmetric nonnormal data condition. Type I error rates for these test statistics are inflated above the 7.5% upper bound of the robustness criterion at all sample size conditions examined. In contrast the k-factor corrected $T_{SB1\text{-}k}$ delivers Type I error rates within the robustness criterion at every sample size investigated, many falling within the 95% CI. As with the elliptical data form, the T_{SB2} and $T_{SB2\text{-}k}$ test statistics appear to over-correct with asymmetric nonnormal data, yielding low model rejection rates (at or near zero for $T_{SB2\text{-}k}$). The unadjusted T_B and T_{ADF} statistics in Table 2 again show near 100% Type I error propensities while the chi-square adjusted forms, T_{B1} and T_{ADF1}, yielded attenuated model rejection rates over the range of sample size conditions investigated. The F statistics again show diverging performance with T_{B2} at sample sizes above its minimum n=df+1 yielding empirical Type I error rates within the robustness criterion. In contrast, model rejection rates associated with T_{ADF2} were generally too small, well below the 2.5% lower limit of the robustness criterion even at the largest sample size condition of n:q=10:1.

The performance of the test statistics under constrained Model A2 is considered next, inspecting results in Table 3. Overall test statistic performance is similar to the behavior found with Model A1, although performance associated with all of the test statistics examined under Model A2 is deteriorated to varying extents. For example with multivariate normal data, Table 3 reports inflated model rejection rates associated with T_{ML} and T_{SB1} that are uniformly higher than those found in Table 2 (again systematically decreasing with increasing sample size). Applying the Bartlett k-factor correction to these test statistics again leads to improved Type I error performance with Model A2 and normal data, although at the smallest sample size conditions $T_{ML\text{-}k}$ and $T_{SB1\text{-}k}$ exert too much correction yielding Type I error rates below the 2.5% lower bound of the robustness criterion. The behavior of T_{SB2} and $T_{SB2\text{-}k}$ is again similar to the unconstrained model; T_{SB2} yields inflated model rejection rates at n:q=1:1 and then operates

nominally at all other sample size conditions while T_{SB2-k} yields near zero model rejection rates across the majority of the sample size conditions investigated. The T_B and T_{ADF} test statistics, and their adjusted forms, again closely mirror one another with respect to Type I error propensities with normal data.

Insert Table 3 about here

Turning now to the elliptically symmetric nonnormal data form, results in Table 3 show that only the T_{SB1-k} test statistic maintained Type I error rates to within the robustness criterion (except for T_{ADF2} at $\underline{n}:\underline{q}$=10:1). Model rejection percentages are inflated for T_{ML}, T_{ML-k}, T_{SB1}, T_B, T_{ADF}, and T_{B2} with elliptically symmetric data while T_{SB2}, T_{SB2-k}, T_{B1}, and T_{ADF1} yield attenuated (and often near zero) Type I error rates. For T_{SB1-k} model rejection rates are within the robustness interval at all but the two smallest sample size conditions (the reported 2.4% Type I error rate at $\underline{n}:\underline{q}$=2:1 falls just below the 2.5% lower bound of the robustness criterion), and in fact are within the 95% CI for the 5% level of significance at these sample size conditions.

Asymmetric nonnormal data (Distribution 3) again has a greater impact on model test statistics than elliptical data. Model test statistics that yielded inflated Type I error rates with elliptical data are even larger with asymmetric data, while model test statistics that were attenuated with elliptical data are further suppressed. Again two notable test statistics yielding robust Type I error rates are T_{ADF2} at its largest sample size condition and T_{SB1-k}. Note, however, the concerning pattern of model rejection rates associated with T_{SB1-k} in Table 3: Type I error rates increase monotonically with increasing sample size, beginning just below the lower bound of robustness at $\underline{n}:\underline{q}$=1:1, falling within the robustness criterion at $\underline{n}:\underline{q}$=2:1 and 5:1, and then exceeding the 7.5% upper bound at all other sample size conditions.

Results for the constrained CFA Model B1 given in Tables 4 are reviewed next. Beginning with Type I error propensities for normal data, Table 4 shows that T_{ML} yielded inflated Type I error rates that decreased with increasing sample size, approaching the 7.5% upper bound of the robustness criterion at

the largest sample size condition. Applying the k-factor correction to yield $T_{ML\text{-}k}$ and $T_{SB1\text{-}k}$ provided improved Type I error control, albeit evidencing some degree of over-correction at the smallest sample size conditions. Also note that with multivariate normal data the T_{SB2} statistic controlled Type I error rates at all but the smallest sample size condition.

Insert Table 4 about here

With nonnormal data, both elliptically symmetric and asymmetric, only $T_{SB1\text{-}k}$ maintained Type I error rates to within the robustness interval. In Table 4 note that T_{ML}, $T_{ML\text{-}k}$, T_{SB1}, T_B, T_{ADF}, T_{B1}, T_{ADF1}, T_{B2}, and T_{ADF2} delivered inflated Type I error rates across all sample sizes investigated (with a few exceptions in which T_{B1}, T_{B2}, and T_{ADF1} yielded 0% model rejections at the smallest sample size conditions). T_{SB2} and $T_{SB2\text{-}k}$ generally yielded low or near zero model rejection percentages under the elliptically symmetric nonnormal data condition. The same pattern of results (although mostly more extreme) is evidenced with asymmetric nonnormal data given in Table 4. Results for the $T_{SB1\text{-}k}$ statistic show a remarkable tendency to control Type I error rates under nonnormal data, even at some of the smallest sample size conditions investigated: Type I error rates associated with $T_{SB1\text{-}k}$ are within the robustness interval (and frequently within the 95% CI) at sample size conditions of n:q=5:1 and greater. At sample size conditions below n:q=5:1 $T_{SB1\text{-}k}$ showed a tendency to over-correct, yielding low model rejection percentages falling just below the 2.5% lower bound of the robustness criterion with extremely nonnormal asymmetric data at n:q=2:1.

The final model investigated is the unconstrained CFA Model B2, with Type I error rates given in Tables 5. With multivariate normal data many of the model fit statistics perform optimally, with model rejection percentages within the defined robustness interval and 95% CI. T_{ML} is robust at the n:q=5:1 and 10:1 sample size conditions; at smaller n:q sample size conditions model rejection rates are inflated. The k-factor corrected forms, $T_{ML\text{-}k}$ and $T_{SB1\text{-}k}$, control inflation in the test statistics with normal data and

deliver Type I error rates that are within the robustness interval at sample size conditions of n:q=1:1 and larger.

Insert Table 5 about here

Examining results for elliptically symmetric data in Table 5 reveals that T_{SB1} controlled Type I error rates to within the robustness interval given adequate sample size-to-parameter ratio (n:q ≥5:1). The k-factor corrected form, T_{SB1-k}, again shows remarkable resistance to test statistic inflation with model rejection rates within the 95% CI at n:q≥1:1. The T_{SB2} and T_{SB2-k} test statistics again tend to deliver low model rejection rates, falling within the robustness interval only at the largest sample size conditions investigated. T_B and T_{ADF} yielded inflated model rejection percentages that monotonically decreased with increasing sample size, approaching the 7.5% upper bound of the robustness criterion. T_{B1} and T_{ADF1} operate nominally at the 5:1 and 10:1 n:q sample size conditions, with low model rejection rates at smaller sample sizes. The F statistics, T_{B2} and T_{ADF2}, showed robustness with respect to Type I error rates across most of the sample size conditions, especially T_{B2} with most of its model rejection percentages falling within the 95% CI.

Test statistic performance under the asymmetric nonnormal data form deteriorates for several of the model fit statistics examined. T_{SB1} controlled test statistic inflation to some extent, but only yielded a model rejection rate falling within the robustness interval at the largest sample size condition investigated. The Bartlett k-factor corrected form, T_{SB1-k}, performed uniformly better than T_{SB1} delivering Type I error rates within the robustness interval at the n:q=5:1 and 10:1 sample size conditions. T_{SB2} and T_{SB2-k} tended to over-correct, as with other models investigated here, yielding low (often near zero) model rejection rates. The behavior of the T_B and T_{ADF} test statistics again mirror one another closely in their unadjusted forms; Type I error rates are inflated and decrease systematically with increasing sample size. The chi-square adjusted forms, T_{B1} and T_{ADF1}, yielded low model rejection rates (with the exception of T_{ADF1} at

$n=p^*$ where it delivers an anomalous rejection rate of 6.45%) while the F statistics here show diverging performance. The residual-based adjusted test statistic T_{B2} is robust at all sample sizes above its $n=df+1$ minimum, while T_{ADF2} yields low model rejection rates, falling below the 2.5% lower bound of the robustness criterion even at the largest sample sizes investigated.

Empirical Power

Results for follow-up analyses examining the power to reject misspecified models are given in Tables 6-9, each table paralleling one of the four models investigated in the Type I error analysis. Power was not analyzed for the residual-based and ADF test statistics at their respective sample size minima due to the high frequency of non-convergence, improper solutions, and test statistic failure rates, and the poor Type I error performance associated with these test statistics at their sample size lower bounds. A general pattern is evidenced across all of the power tables in which power systematically increased with increasing sample size (as would be expected) and uniformly decreased with increasing departure from multivariate normality.

Insert Table 6 about here

Table 6 presents empirical power estimates for the misspecified Model A1. At sample sizes less than $n=350$ test statistic performance with respect to the power to reject a misspecified model diverges, allowing for some noteworthy comparisons. The Bartlett k-factor correction to T_{ML}, $T_{ML\text{-}k}$, remained powerful down to the smallest sample size condition (under multivariate normal and elliptically symmetric data) with empirical power reported at 81-84%. Note the power associated with $T_{SB1\text{-}k}$ which is reported at 100% under all distributional forms down to the $n=70$ sample size condition, and near 90% at the $n=35$ sample size. In contrast, the T_{SB2} and $T_{SB2\text{-}k}$ test statistics delivered low power at the smaller

sample size conditions – power falls to 50% for T_{SB2} with asymmetric nonnormal data and to only about 5% for $T_{SB2\text{-}k}$ at the n=35 sample size.

Empirical power estimates for the residual-based and ADF test statistics in Table 6 allow for several different comparisons. First, comparing T_{B1} and T_{B2} against T_{ADF1} and T_{ADF2} (at the n=175 sample size) note that the residual-based test statistics generally exhibited greater power to reject the misspecified model than the ADF test statistics. Also, the F-corrected forms of T_B and T_{ADF} (T_{B2} and T_{ADF2}) delivered greater power as compared to the chi-square-corrected forms of these test statistics (T_{B1} and T_{ADF1}). This pattern of results is seen comparing T_{B1} against T_{B2} at both the n=120 and n=175 sample size conditions (note T_{B1} at n=120 has zero or near zero power estimates), and comparing T_{ADF1} against T_{ADF2} at the n=175 sample size.

Turning now to misspecified Model A2, Table 7 reports empirical power estimates that are noticeably lower than those found in Table 6. For the $T_{ML\text{-}k}$ test statistic empirical power is reported at 20% at the smallest sample size condition for multivariate normal data, with power at or near 100% at all larger sample sizes. The k-factor corrected $T_{SB1\text{-}k}$ reports power at or near 100% at all $n \geq 90$; empirical power is attenuated at n=18 with all data forms (less than 30% with nonnormal data), and degraded to below 50% at n=36 with the more severe asymmetric nonnormal data. The T_{SB2} and $T_{SB2\text{-}k}$ test statistics yielded 90-100% power under normal and elliptically symmetric data across the larger sample sizes investigated ($n \geq 90$). However, the power associated with these test statistics tends to be low (less than 50%) with asymmetric nonnormal data at all but the largest sample size condition.

Insert Table 7 about here

The residual-based and ADF test statistics are analyzed with respect to power under only a few conditions with misspecified Model A2. For this model comparing T_{B1} against T_{ADF1} in Table 7 shows equivalent test statistic performance and generally low power with nonnormal data forms. Again notice

the 0% model rejection rate under all data forms for T_{B1} at the $\underline{n}$=120 sample size. Comparing the chi-square-corrected T_{ADF1} against the $\underline{F}$ statistic T_{ADF2} again indicates greater power associated with the $\underline{F}$ statistic.

Empirical power estimates for misspecified Model B1 given in Table 8 are considered next. Power estimates are quite high for T_{ML-k} at all sample sizes, dipping below 100% only at the smallest sample size of $\underline{n}$=35. The T_{SB1-k} test statistic also showed a strong ability to reject the misspecified model with power estimates reported at 90-100% over most of the study conditions examined; power does however become degraded at the smallest sample size conditions with nonnormal data. The T_{SB2} and T_{SB2-k} test statistics generally yielded high empirical power with normal and elliptically symmetric data forms with power becoming noticeably attenuated with asymmetric nonnormal data, and also at the $\underline{n}$=35 sample size. Little can be said regarding the performance of the residual-based and ADF test statistics here as these model fit statistics were analyzed under only a few cells for misspecified Model B1. Again, notice that the chi-square test statistic T_{B1} yields 0% power under all data forms at the $\underline{n}$=231 sample size (a sample size close to is minimum lower bound of $\underline{n}$=218).

Insert Table 8 about here

Power estimates in Table 9 for misspecified Model B2 show a predominance of 100% rejection rates for most of the test statistics analyzed. Most noteworthy is the performance of the adjusted residual-based (T_{B1} and T_{B2}) and ADF (T_{ADF1} and T_{ADF2}) test statistics at the small and moderate sample sizes. These test statistics again exhibit the previously noted pattern with respect to the power to reject the misspecified model; the residual-based test statistics are uniformly more powerful than their ADF analogs, with the $\underline{F}$ statistics T_{B2} and T_{ADF2} delivering greater power than the chi-square-corrected T_{B1} and T_{ADF1} test statistics. Also, note that for the T_{B2} test statistic at the $\underline{n}$=129 sample size (under all three

distributional forms) its corresponding Type I error rates where held to within the robustness criterion, however, the observed power estimate associated with T_{B2} at this sample size is quite low, at 16-26%.

Insert Table 9 about here

Discussion

Results in the present investigation replicate previous findings and extend into new directions our understanding of assessing global model fit in SEM. This section reflects upon several compelling issues that necessitated such an investigation. First, test statistic performance across the models under investigation is considered, assessing the impact of model type, size, and complexity on test statistic performance. Second, study results are reviewed from a perspective of assessing sample size adequacy. Next, the main findings are briefly summarized for each of the test statistics investigated with recommendations proffered to practitioners. Some final remarks conclude this report.

Model Type, Size, and Complexity: Impact on Test Statistic Performance

The four models examined in the present investigation were parameterized specifically to address issues relating model type, size, and complexity to test statistic performance. To begin, test statistic performance under the unconstrained 85 df Model A1 in this study is compared against results reported by Yuan and Bentler (1998, 1999) and by Bentler and Yuan (1999). The model under scrutiny in those studies was a 5-factor 15-variable CFA model requiring q=33 parameters to be estimated and yielding 87 df, which is reasonably equivalent in size and complexity to Model A1 here but varying in model type (CFA vs. SE). Test statistic performance for T_{ML}, T_{SB1}, the residual-based, and the ADF test statistics in the present study for Model A1 is remarkably similar to the reported findings under the previously

BEST COPY AVAILABLE

examined CFA model with respect to both patterns of results and absolute model rejection rates. Such consonance suggests that model type has little to no apparent influence on test statistics' Type I error propensity or on the power to reject misspecified models.

Another issue that can be addressed using the results of the present investigation is test statistic performance in larger variable systems. Previous investigations have demonstrated the potential effectiveness of the test statistics examined here with smaller models using p=6 or 15 variables and few latent factors (for example k=0 or 3). The present investigation is the first to examine a generally larger latent variable system with p=21 variables and k=7 latent factors. Results for the unconstrained Model B2 bode well for assessing model fit in SEM with larger latent model structures. With multivariate normal data the T_{ML-k}, T_{SB1-k}, T_{SB2}, T_{B2}, and T_{ADF2} test statistics controlled Type I error rates at the smaller sample size conditions investigated while maintaining strong power to reject the misspecified model. Strong test statistic performance under this model with nonnormal data further indicates that larger latent systems do not necessarily disrupt the ability to assess global model fit.

Another perplexing issue, addressed by Yuan and Bentler (1998), is the impact of increasing model df on test statistic performance. Yuan and Bentler (1998) compared test statistic performance between the 87 df CFA model and a constrained 118 df model (with q=2 parameters estimated). They reported that T_{SB1} and T_{B1} exhibited some degradation in performance under the constrained 118 df model, while T_B and its adjusted F statistic T_{B2} were severely distorted under the constrained model. They concluded that the residual-based test statistic T_B is particularly sensitive to large df, rather than model complexity (as measured by the q), and that such distortion is propagated into its adjusted form T_{B2}.

By examining results in the present investigation the issue of increasing df and its impact on test statistics may be further advanced. Contrasting test statistic performance under the unconstrained and constrained models here leads to similar conclusions and new findings. First, model non-convergence

and improper solution rates for both ML and ADF estimation were lower with constrained (larger df) models as compared to their respective unconstrained forms. However, the constrained models also evidenced increased rates of test statistic failure for the residual-based test statistic (i.e., T_B more frequently inflated to values exceeding the arbitrary cut-off of 900,000 used to define residual-based test statistic failure). This finding for T_B is in agreement with the inflation in the residual-based test statistic under the larger df model as reported by Yuan and Bentler (1998).

In the present investigation absolute sample sizes differed across models in order to preserve n:q ratios, thus making direct comparisons to the findings by Yuan and Bentler (1998) somewhat difficult. To maintain equivalence with that study, comparisons across models in the present investigation focus on sample size conditions in which absolute sample sizes are equal (or at least close to equal). Evidence here suggests that test statistics based on T_{ML} are only minimally impacted by model parameterization, while the residual-based test statistic T_{B2} is more prominently affected. For example, compare test statistic performance in Models A1 and A2 at the common sample size of n=120. With nonnormal data adjusted test statistics based on T_{ML} appear to be stable across models, while T_{B2} yields inflated Type I error rates under the larger df model.

The design of present study allows for further investigation into the failure of the residual-based T_{B2} detailed above. If increasing df is the proximal cause of such test statistic failure, as has been conjectured (Yuan & Bentler, 1998), then one would expect comparable performance of T_{B2} under the constrained Model A2 and unconstrained Model B2, both yielding 102 df. With normal data and at a comparable absolute sample size of n=120 under Model A2, and n=129 under Model B2, the T_{ML} based test statistics yield Type I error rates that are very similar across models, while T_{B2} is inflated under Model A2 but robust under Model B2. With nonnormal data the T_{SB1} and $T_{SB1\text{-}k}$ test statistics again remained reasonably stable across models while the T_{B2} test statistic was inflated under the constrained Model A2, but was again robust under the unconstrained Model B2. Such evidence contradicts increasing

model df as a proximal cause, suggesting instead that an artifact of constraining models to yield larger df may be at issue.

The argument implicating model parameterization rather than increasing model df for the poor performance of T_{B2} is further bolstered by evaluating the F statistic under the unconstrained Model A1 (85 df) against the unconstrained Model B2 (102 df). Again a common sample size across the two models is used to contrast Type I error rates associated with the test statistic with nonnormal data (n=120 for Model A1, n=129 for Model B2). Under these models note that Type I error propensity associated with T_{B2} is remarkably similar; Type I error rates under both models for the F statistic are within the robustness criterion at about 6.5%. Thus, here is evidence that increasing model df has no apparent impact on the performance of the T_{B2} test statistic when unconstrained models are examined.

In sum, results in the present investigation are consistent with those of Yuan and Bentler (1998) in that the performance of the T_{B2} test statistic became degraded under varying model conditions. However, evidence in the present study suggests that their conclusion implicating increasing model df may be suspect, with a more likely cause for inflation in the test statistic attributable to constraints imposed upon the model. Regardless of the cause, findings in the present investigation do not bode well for employing the residual-based F statistic as a method for assessing global model fit in SEM. Clearly a test statistic that behaves erratically – operating correctly under some model parameterizations but becoming distorted under other model specifications – should not be embraced by practitioners as a reliable fit statistic.

Assessing Sample Size Adequacy

Since Tanaka (1987) conjectured that sample size adequacy in SEM might best be gauged using an n:q ratio rather than using absolute sample size, the methodological community has embraced this approach for providing sample size guidelines to practitioners. Results in the present investigation support the notion that this ratio is important in assessing sample size adequacy. However, Type I error

results in this investigation raise new questions regarding the adequacy of basing sample size requirements solely on the ratio of n:q, and open the door to further examination.

For example, consider the Type I error propensity of T_{ML} evidenced in the present investigation with multivariate normal data (the following discussion is also applicable to $T_{ML\text{-}k}$ with normal data, and to the T_{SB1} and $T_{SB1\text{-}k}$ test statistics with nonnormal data forms). Comparing model rejection rates for, say, Models A1 and A2, notice that model rejection rates appear to be more comparable across models with respect to absolute sample sizes rather than n:q ratio. For Models B1 and B2, despite the fact that absolute sample sizes do not match up as closely in these models as they do under Models A1 and A2, the pattern of Type I error behavior more closely paralleling absolute sample size rather than n:q ratio prevails.

At this point evidence in the present investigation suggests that, at least in part, absolute sample size may be driving Type I error propensity. Still, consider results (again for T_{ML} with normal data) under Models A1 and B1 in which four of the sample size conditions are equivalent with respect to both n and n:q (n=35, 70, 175, and 350). Model rejection rates across sample sizes in Table 6 are lower than those in Table 12, most notably at the smallest sample sizes of n=35 and 70. Such findings are suggestive that characteristics beyond absolute sample size and n:q ratio may play an important role with respect to sample size adequacy.

One potentially important aspect to test statistic performance may be the size of the variance-covariance matrix modeled (i.e., p*). Notice that Model A1 requires the sample data to estimate p*=120 variances and covariances while under Model B1 p*=231 variances and covariances are estimated. Models A1 and B1 do vary in model df as well as p*, however, the previous discussion touched on the effect of imposing constraints upon the model, concluding that such constraints have little impact on test statistics based on T_{ML}. Still, Models A2 and B2 hold model df constant; contrasting these models may be useful in considering this issue further. Examining the performance of T_{ML} with multivariate normal

data under these models shows that at comparable absolute sample sizes (n=103, n=120 and 129) model rejection rates are quite close. Unfortunately, Models A2 and B2 differ in both p^* and q, thus their comparison is not without confound.

In an attempt to clarify the issue of gauging sample size adequacy in SEM, a post hoc exploratory examination of the relations between empirical Type I error rates and design characteristics was conducted. The relation between Type I error rates and p^*, q, df, k, n, and n/q was inspected; also included in the exploratory analysis were previously unconsidered (but potentially informative) characteristics of n/p^*, n/df, n/k, and n-df. The analysis was conducted by tabling columns for empirical error rate and the design characteristics; rows of the table were then sorted by ascending error rate (these tables are available from the author upon request). Design characteristics were inspected to assess which characteristics tended to sort with error rate; to aid in this inspection Pearson correlations were computed between error rate and the various design characteristics.

T_{ML} with multivariate normal data was examined first. The variable most strongly related to Type I error rate was the n:q ratio with a correlation of -.499, while the n:p^* ratio was just slightly less strongly related to error rates with a correlation of -.433. Rows in the table reflecting inflated Type I error rates (i.e., falling outside the previously defined robustness interval) exhibit n:q ratios that are generally less than 10:1, and interestingly have n:p^* ratios that are mostly below 1.5:1. For the inflated Type I error rates there are several values of n:q that equal or exceed the often recommended minimum value of 5:1, but for these cases the corresponding n:p^* ratio was nearly always less than 1.5:1.

Because the k-factor corrected form of T_{ML} showed robustness with normal data at sample sizes smaller than those for which T_{ML} was robust, Type I error rates were examined for $T_{ML\text{-}k}$. In this case there are strong positive correlations with error rate and n, n:p^*, n:q, n:df, n:k, and n-df. Such findings are not surprising for $T_{ML\text{-}k}$ as the k-factor correction incorporates p, k, and n into the adjustment to T_{ML}. Still, as with the examination of T_{ML}, the n:p^* and n:q ratios exhibit the strongest linear relation with

Type I error rates with correlations of .728 and .693, respectively. Based on n:q ratios, it might appear that an n:q=2:1 ratio is sufficient to control error rates to within the robustness interval. However, an additional measure may be in n:p*; while some n:q=2:1 ratios fall outside the robustness interval, all n:p* ratios of .5:1 and larger are associated with robustness in the T_{ML-k} test statistic. Thus the k-factor correction to T_{ML} may reduce the requisite n:p* ratio from 1.5:1 down to .5:1.

In the present investigation the k-factor corrected T_{SB1} test statistic evidenced the most stable performance, controlling Type I error rates at small sample sizes with normal and nonnormal data forms and across the varying model parameterizations. To assess sample size issues associated with this test statistic correlation analyses with error rate and design characteristics were conducted and, similar tables were constructed for each of the three distributional forms. Results for the T_{SB1-k} test statistic were mixed, but again suggest that a combination of n:q≥2:1 and n:p* ≥.5:1 appear to be governing lower bound ratios.

Clearly there is evidence in the present investigation, albeit rather limited, suggesting that the n:p* ratio may be important for assessing sample size adequacy in SEM , possibly in conjunction with the n:q ratio. It is not unreasonable to conjecture that both ratios might play important roles in determining minimum sample sizes in SEM. It has been long established that the number of model parameters requiring estimation should factor into sample size issues. Still, results in the present investigation suggest that considering q in the absence of accounting for the number of sample variances and covariances that require estimation may not necessarily assure successful modeling. From the examinations here it appears that both ratios should be considered together to determine adequate minimum sample sizes. Clearly findings in the present study addressing the issue of assessing sample size adequacy in SEM are tenuous at best; the present investigation was not designed to specifically address this issue and as a result q and p* characteristics here are confounded. Further investigation, systematically manipulating n:q and n:p* ratios, is required to more fully profile the potential importance of the n:p* ratio in assessing sample size adequacy and its relation to the n:q ratio.

BEST COPY AVAILABLE

Summarizing Test Statistic Performance and Recommendations to Practitioners

With respect to both model estimation failure rates using ML and the Type I error propensity for T_{ML}, results in the present investigation closely follow previous findings. Beginning with rates of non-convergence and improper solutions, results in the present investigation lead to a caution that modeling data with sample size ratios at $\underline{n}:q \leq 1:1$ is unwise. While non-convergence was not observed to be problematic, rates of improper solutions at this smallest sample size ratio were often high, especially with nonnormal data forms; it is unlikely that practitioners will be able to obtain usable model estimates at sample size ratios of $\underline{n}:q=1:1$ or smaller.

Considering model rejection behavior, T_{ML} was not robust to departures from multivariate normality in the present investigation, yielding inflated Type I error rates even at the largest sample sizes or sample size-to-estimated parameter ratios examined. With normal data, T_{ML} also was inflated under most of the sample size conditions examined, most notably often at $\underline{n}:q=5:1$ and 10:1, ratios that meet or exceed the guideline for using ML estimation that has become pervasive in the SEM community. Such results cast doubt on such recommended guidelines for T_{ML}.

Findings in the present study associated with the SB test statistics were mixed, with Type I error rates associated with the SB scaled T_{SB1} statistic inflated at most $\underline{n}:q<10:1$. Type I error rates corresponding to the SB adjusted T_{SB2} test statistic were mostly attenuated under nonnormal data conditions and power was frequently low. Based on such results, use of the T_{SB2} test statistic is not warranted. The test statistic was developed to accommodate the inflation observed in T_{ML} with nonnormal data forms, however, here the test statistic performed optimally only with multivariate normal data.

The Type I error propensity and power associated with the Bartlett $\underline{k}$-factor corrected forms of T_{ML}, T_{SB1} and T_{SB2} in the present investigation are reviewed next. Applying the $\underline{k}$-factor correction to T_{SB2} in this study was not especially effective –adjusting T_{SB2} with the $\underline{k}$-factor correction further reduced

an already attenuated test statistic. Unlike the performance of T_{SB2-k}, the T_{ML-k} and T_{SB1-k} fit statistics generally maintained Type I error rates to within the robustness interval at small sample sizes with normal (T_{ML-k} and T_{SB1-k}) and nonnormal data forms (T_{SB1-k}). The performance of T_{SB1-k} in the present investigation is particularly noteworthy, with the combination of SB scaling and $\underline{k}$-factor correction yielding the most stable test statistic investigated here. Type I error rates associated with T_{SB1-k} were close to the 5% level at nearly all conditions investigated and, the test statistic exhibited high power under all but the smallest sample sizes examined. Based on such findings, practitioners are encouraged to turn to this test statistic to assess global model fit when sample sizes are small to moderate.

It should be noted that in the power analysis models were misspecified with respect to the number of latent factors ($\underline{k}$), in part to assess the performance of the $\underline{k}$-factor correction under conditions in which the value of $\underline{k}$ used in the adjustment did not equal the true number of latent factors in the population. For all misspecified models, $\underline{k}$ was underestimated. Models were misspecified using fewer factors than were present in the true population model to mirror (parsimonious) modeling behavior exhibited by practitioners. Recall that the multiplicative correction factor applied either to T_{ML} or to the SB test statistics is of the form $\underline{c}=1-(2\underline{p}+4\underline{k}+5)/6(\underline{n}-1)$. Setting $\underline{k}$ to a value that was smaller than the true population value for the misspecified models in this study yielded a multiplicative constant that was larger than would be obtained had the correct value of $\underline{k}$ been used. The net effect being that the $\underline{k}$-factor adjustment pushed the corrected test statistics to larger values, making it easier to reject the false null hypothesis. Thus for the power analysis the forms of misspecification in the present investigation may have afforded the $\underline{k}$-factor adjustment some degree of advantage. Nevertheless conditions in the present investigation do represent what might be evidenced by practitioners, and in that sense results here are representative of the behavior of the $\underline{k}$-factor corrected test statistics with misspecified models.

Results in the present study (and others) suggest that the SB scaling to T_{ML} delivers a test statistic (T_{SB1}) that is not distorted by nonnormal data forms, but does require medium to large sample sizes to

perform optimally. In tandem to such findings, this investigation and the Fouladi (1998) study have shown that the k-factor correction to T_{ML} with multivariate normal data can control inflation in the test statistic at small sample sizes. Thus it is not unreasonable that the k-factor correction to T_{SB1} might perform so consistently under a variety of study conditions – the SB scaling controls inflation in the test statistic brought on by nonnormality and the k-factor correction controls inflation due to small sample size. This dual combination of corrections to T_{ML} may be the most potent and stable strategy to obtaining reliable model fit statistics in SEM. Certainly evidence supporting the effectiveness of the k-factor correction is limited, and more investigation applying the k-factor correction to the SB test statistic is warranted. Specifically follow-up studies systematically manipulating k, particularly with misspecified models where the number of latent factors might be over-specified as well as under-specified should be informative to practitioners and methodologists alike.

The performance of $T_{SB1\text{-}k}$ not withstanding, the residual-based and ADF test statistics have received much attention in the literature and were thus examined in the present investigation. Type I error rates observed in the present study for T_B and T_{ADF} have again shown that these test statistics are not useful for assessing model fit in SEM at realistic sample sizes. The chi-square and F-corrected forms of the residual-based and ADF statistics showed some promise in this investigation, maintaining Type I error rates to within the defined robustness criterion and delivering strong power to reject misspecified models. Performance of these corrected test statistics was erratic, however, controlling Type I error rates under some conditions and failing under others. Moreover, the T_{B1}, T_{ADF1}, and T_{ADF2} test statistics behaved unpredictably when they failed, becoming inflated under some study conditions and becoming attenuated under others.

The present investigation provided the opportunity to directly compare the analogous forms of the residual-based and ADF test statistics against one another under a common set of experimental conditions. The unadjusted test statistics T_B and T_{ADF} mirrored one another closely in Type I error

propensity throughout the present study, as did the chi-square-adjusted forms, T_{B1} and T_{ADF1}. With respect to the power to reject misspecified models, the chi-square statistics were close to one another in performance, with the residual-based T_{B1} statistic evidencing somewhat greater power than T_{ADF1} under unconstrained models with nonnormal data and at small sample sizes. Unlike the chi-square statistics, the F statistics T_{B2} and T_{ADF2} paralleled one another under some study conditions and diverged from one another under other conditions. Under the unconstrained models with nonnormal data (at sample sizes above their respective minima) T_{B2} showed better performance than T_{ADF2}; T_{B2} tended to control Type I error rates while T_{ADF2} yielded attenuated error rates. Not surprisingly with misspecified models under conditions in which T_{ADF2} yielded suppressed Type I error rates, T_{B2} exhibited better power than T_{ADF2}.

Overall, results in the present investigation favor an endorsement of the residual-based test statistics over their analogous ADF fit measures. First, non-convergence and improper solution rates favor ML estimation (from which the residual-based test statistics were derived) over ADF estimation. Second, minimum sample size constraints also promote the residual-based statistics over those stemming from ADF estimation, as $df+1$ will nearly always be less than p^* (whenever $q>1$). Finally based on Type I error propensity and power, findings in the present investigation again favor the residual-based test statistics, T_{B2} in particular which may be the most useful model fit statistic in this class of test statistics. Certainly practitioners should exercise caution using this test statistic at sample sizes close to its lower bound $n=df+1$.

Closing Remarks

Based on findings in the present investigation it appears that practitioners with small to moderate sample sizes can effectively model their data and accurately assess global model fit using SEM techniques. Even under severely nonnormal data conditions some of the test statistics investigated here yielded strong performance at sample sizes less than n=100 observations, some fit statistics even operated well at less than n=50 subjects. Certainly such results bode well for SEM use with smaller sample sizes, but other results in the present investigation (in which test statistics failed) also indicate that such modeling techniques require careful attention to particular design characteristics. Throughout the present study the number of model parameters requiring estimation (q) and the size of the variance-covariance matrix applied to the models (p*) played important roles in test statistic performance and thus cannot be ignored by practitioners applying SEM techniques to their data.

New findings in the present investigation support the use of the Bartlett k-factor correction for controlling test statistic performance at small sample sizes, particularly when combined with the SB scaled form of T_{ML} under nonnormal data conditions. Type I error rates were controlled and power was generally high using this test statistic with both normal and nonnormal data forms and across most of the model specifications in the present investigation. The importance of assessing sample size adequacy with regard to the size of the sample variance-covariance matrix as well as with respect to the number of freely estimated model parameters has also come to light through the present study. As stated previously, follow-up studies that systematically vary these two seemingly important ratios could better define minimum requisite values for accurate modeling.

Recommendations to practitioners for successful covariance structure modeling with small samples are as follows. First, ML is the recommended estimation method of choice; ML does not hold the n=p* minimum sample size restriction that is implicit in ADF estimation, thus allowing for modeling at smaller sample sizes. To protect against potential problems with model non-convergence or improper

solutions using ML, it is recommended that practitioners obtain sample sizes large enough such that the sample size-to-estimated parameter ratio is no less than n:q=2:1. Evidence here leads to an endorsement of the k-factor-corrected SB scaled test statistic for assessing overall model fit with small samples. Practitioners are recommended to turn to this statistic in their applied modeling and need not be overly concerned with meeting distributional assumptions or with the parameterization of their models. Moreover, it is a test statistic that is readily available to practitioners; the unadjusted SB scaled test statistic is available in SEM programs such as LISREL and EQS, with the k-factor-corrected form easily computed and assessed for statistical significance. Finally, for the prudent use of this test statistic, it is recommended that practitioners should also account for the ratio of sample size-to-number of sample variances and covariances in concert with the n:q ratio; findings here recommend that a combination of n:q≥2:1 and n:p*≥.5:1 should be sufficient for the effective use of the k-factor-corrected SB scaled test in applied modeling scenarios.

The present investigation has focused on assessing global model fit in SEM using inferential test statistics at small samples. While ground has been gained in understanding assessing overall model fit at reduced sample sizes, other issues remain largely uninvestigated. One issue is the impact of model test statistics (such as T_{ML}) on incremental and absolute model fit indices at small sample sizes and with nonnormal data. The present investigation has shown that applying corrective measures to T_{ML} such as the k-factor correction and SB scaling are useful for counteracting the negative influences on the model test statistic due to small sample size and nonnormality. A natural extension to these corrective measures is to incorporate such adjusted test statistics into other indices of model fit. In fact, recent releases of popular SEM software now offer some robust fit indices; EQS 6.0 will provide a "robust" comparative fit index (CFI) which is computed using T_{SB} rather than T_{ML}. While some research has been conducted examining the behavior of such modified fit indices (Anderson, 1996; Nevitt & Hancock, in press-a), with promising results, there is clearly the need for further examination of modified fit indices that might be

BEST COPY AVAILABLE

corrected for small samples as well as nonnormal data forms, incorporating, for example, the T_{SB1-K} test statistic.

Another issue warranting further investigation is the quality of individual estimated model parameters at small sample sizes. While global model fit may be adequately assessed at small sample sizes, it remains to be seen how well individual model parameters can be estimated at such reduced sample size conditions. Moreover, accurately assessing the statistical significance of individual model parameters may be tenuous with small samples; parameter standard errors may not necessarily be stable under such conditions leading to distorted z-statistics and incorrect inference. Still, through continued research such as the present investigation SEM techniques continue to evolve, breaking sample size and distributional barriers, thus becoming more accessible and useful to practitioners.

REFERENCES

Anderson, R. D. (1996). An evaluation of the Satorra-Bentler distributional misspecification correction applied to the McDonald fit index. Structural Equation Modeling: A Multidisciplinary Journal, 3, 203-227.

Aptech Systems. (1996). GAUSS System and Graphics Manual. Aptech Systems, Inc, Maple Valley, WA.

Arbuckle, J. L. (1997). AMOS Users' Guide, Version 3.6. Chicago, IL: SmallWaters.

Bartlett, M. S. (1950). Tests of significance in factor analysis. British Journal of Psychology, Statistical Section, 3, 77-85.

Bentler, P. M. (1996). EQS Structural Equations Program Manual. Encino, CA: Multivariate Software, Inc.

Bentler, P. M., & Chou, C.-P. (1987). Practical issues in structural modeling. Sociological Methods & Research, 16, 78-117.

Bentler, P. M., & Yuan, K.-H. (1999). Structural equation modeling with small samples: Test statistics. Multivariate Behavioral Research, 34, 181-197.

Boomsma, A. (1982). The robustness of LISREL against small sample sizes in factor analysis models. In K. G. Jöreskog & H. Wold (Eds.), Systems under indirect observation: Causality, structure, prediction (Part 1). Amsterdam: North-Holland.

Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in LISREL maximum likelihood estimation. Psychometrika, 50, 229-242.

Bradley, J. V. (1978). Robustness? British Journal of Mathematical and Statistical Psychology, 31, 144-152.

Browne, M. W. (1982). Covariance structures. In D. M. Hawkins (Ed.), Topics in applied multivariate analysis (pp. 72-141). Cambridge, England. Cambridge University Press.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. British Journal of Mathematical and Statistical Psychology, 37, 62-83.

Chou, C.-P., Bentler, P. M., & Satorra, A. (1991). Scaled test statistics and robust standard errors for non-normal data in covariance structure analysis: A Monte Carlo study. British Journal of Mathematical and Statistical Psychology, 44, 347-357.

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. Psychological Methods, 1, 16-29.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. Psychometrika, 43, 521-532.

Fouladi, R. T. (1998, April). Covariance structure analysis techniques under conditions of multivariate normality and nonnormality – modified and bootstrap based test statistics. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Fouladi, R. T. (1999, April). Model fit in covariance structure analysis under small sample conditions – Modified maximum likelihood and asymptotically distribution free generalized least squares procedures. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Gerbing, D. W., & Anderson, J. C. (1985). The effects of sampling error and model characteristics on parameter estimation for maximum likelihood confirmatory factor analysis. Multivariate Behavioral Research, 20, 255-271.

Hayduk, L. A. (1987). Structural equation modeling with LISREL: Essentials and advances. Baltimore, MD: The Johns Hopkins University Press.

Hu, L.-T., Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), Structural equation modeling: Concepts, issues and applications (pp. 76-99). Thousand Oaks, CA: Sage.

Hu, L.-T., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? Psychological Bulletin, 112, 351-362.

Jöreskog, K. G., & Sörbom, D. (1996). LISREL 8: User's reference guide. Chicago, IL: Scientific Software International.

Nevitt, J., & Hancock, G. R. (in press-a). Improving the root mean square error of approximation for nonnormal conditions in structural equation modeling. Journal of Experimental Education.

Nevitt, J., & Hancock, G. R. (in press-b). Performance of bootstrapping approaches to model test statistics and parameter standard error estimation in structural equation modeling. Structural Equation Modeling: A Multidisciplinary Journal.

Nevitt, J., & Hancock, G. R. (1999). PWRCOEFF & NNORMULT: A set of programs for simulating multivariate nonnormal data. Applied Psychological Measurement, 23, 54.

Satorra, A., & Bentler, P. M. (1988). Scaling corrections for chi-square statistics in covariance structure analysis. American Statistical Association 1988 proceedings of the Business and Economics Sections (pp. 308-313). Alexandria, VA: American Statistical Association.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), Latent variables analysis: Applications for developmental research (pp.399-419). Thousand Oaks, CA: Sage.

Tanaka, J. S. (1984). Some results on the estimation of covariance structure models. Dissertation Abstracts International, 45, 924B.

Tanaka, J. S. (1987). "How big is enough?": Sample size and goodness of fit in structural equation models with latent variables. Child Development, 58, 134-146.

Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. Psychometrika, 48, 465-471.

Yuan, K.-H., & Bentler, P. M. (1997). Mean and covariance structure analysis: Theoretical and practical improvements. Journal of the American Statistical Association, 92, 767-774.

Yuan, K.-H., & Bentler, P. M. (1998). Normal theory based test statistics in structural equation modeling. British Journal of Mathematical and Statistical Psychology, 51, 289-309.

Yuan, K.-H., & Bentler, P. M. (1999). F tests for mean and covariance structure analysis. Journal of Educational and Behavioral Statistics, 24, 225-243.

Table 1. Numbers of Variables, Unique Variances/Covariances, Estimated Parameters, Degrees of Freedom, and Absolute Sample Sizes for Properly Specified Models

					Sample Size Condition					
Model	p	p*	q	df	1:1	2:1	5:1	10:1	df+1	p*
					n:					
A1	15	120	35	85	35	70	175‡	350‡	86†	120‡
A2	15	120	18	102	18	36	90	180‡	103†	120‡
B1	21	231	35	196	35	70	175	350‡	197†	231‡
B2	21	231	129	102	129†	258‡	645‡	1290‡	103†	231‡

Note. p=number of variables; p*=number of unique variances/covariances; q=number of freely estimated model parameters; df=degrees of freedom. Sample sizes flagged with † are sufficiently large to compute the residual-based test statistics. Sample sizes flagged with ‡ are sufficiently large to compute residual-based and ADF test statistics.

Table 2. Type I Error Rates for Properly Specified Model A1 (85 df)

n-cond:		1:1	2:1	5:1	10:1	df+1	p*
n:		35	70	175	350	86	120
	D						
T_{ML}	1	44.0	17.0	9.65	6.05†	14.0	10.0
	2	48.0	24.0	14.0	10.0	19.0	16.0
	3	75.0	63.0	62.0	70.0	62.0	61.0
T_{ML-k}	1	1.10	2.30	4.45‡	3.90†	2.90†	3.10†
	2	1.70	4.95‡	7.70	7.50†	5.50‡	5.95‡
	3	18.0	37.0	52.0	65.0	39.0	47.0
T_{SB1}	1	59.0	23.0	11.0	6.40†	17.0	12.0
	2	64.0	26.0	11.0	6.65†	18.0	14.0
	3	78.0	34.0	13.0	8.20	26.0	18.0
T_{SB1-k}	1	2.85†	3.75†	5.10‡	4.25‡	4.15‡	3.70†
	2	2.70†	4.75‡	4.75‡	4.40‡	4.75‡	4.55‡
	3	5.05‡	6.80†	5.95‡	5.35‡	6.65†	6.75†
T_{SB2}	1	10.0	3.90†	4.45‡	3.70†	3.75†	2.85†
	2	5.15‡	1.60	1.35	1.55	1.55	1.35
	3	2.50†	0.50	0.20	0.45	0.30	0.10
T_{SB2-k}	1	0.005	0.45	1.40	2.70†	0.50	1.10
	2	0.0	0.0	0.30	1.15	0.30	0.25
	3	0.0	0.0	0.0	0.20	0.05	0.0
T_B	1			99.0	67.0	100.0	100.0
	2			99.0	66.0	100.0	100.0
	3			99.0	65.0	100.0	100.0
T_{B1}	1			1.85	4.60‡	0.0	0.0
	2			0.85	3.15†	0.0	0.0
	3			0.30	1.20	0.0	0.0
T_{B2}	1			12.0	8.50	0.0	12.0
	2			6.15†	5.70‡	0.0	6.50†
	3			3.85†	2.95†	0.0	6.60†
T_{ADF}	1			99.0	67.0		100.0
	2			99.0	64.0		100.0
	3			99.0	60.0		99.0
T_{ADF1}	1			0.65	4.75‡		0.0
	2			0.0	2.65†		0.05
	3			0.0	0.70		0.20
T_{ADF2}	1			6.95†	8.65		0.50
	2			1.60	4.95‡		0.20
	3			0.20	1.60		0.60

Note. Type I error rates are computed across 2000 replications at the 5% level of significance. D=distributional form; distribution 1=multivariate normal; distribution 2=elliptically symmetric nonnormal; distribution 3=asymmetric nonnormal. † indicates rejection rate within the robustness interval; ‡ indicates rejection rate within the 95% CI.

Table 3. Type I Error Rates for Properly Specified Model A2 (102 df)

n-cond:		1:1	2:1	5:1	10:1	df+i	p*
n:		18	36	90	180	103	120
	D						
T_{ML}	1	99.0	43.0	14.0	9.20	13.0	12.0
	2	100.0	80.0	62.0	59.0	63.0	63.0
	3	100.0	98.0	99.0	99.0	99.0	99.0
T_{ML-k}	1	0.05	0.85	2.85†	4.50‡	3.55†	3.20†
	2	1.30	13.0	37.0	47.0	41.0	43.0
	3	20.0	75.0	95.0	99.0	96.0	97.0
T_{SB1}	1	100.0	59.0	18.0	11.0	16.0	14.0
	2	100.0	61.0	18.0	11.0	18.0	15.0
	3	100.0	68.0	22.0	15.0	20.0	17.0
T_{SB1-k}	1	1.65	3.0†	3.65†	5.50‡	4.65‡	4.45‡
	2	1.50	2.40	4.95‡	5.70‡	4.75‡	5.55‡
	3	2.30	4.40‡	6.70†	8.45	7.95	7.55
T_{SB2}	1	72.0	7.0†	2.60†	3.55†	3.15†	2.75†
	2	53.0	1.65	0.55	1.15	0.50	0.95
	3	34.0	0.50	0.30	0.35	0.25	0.20
T_{SB2-k}	1	0.0	0.0	0.40	1.60	0.35	0.80
	2	0.0	0.0	0.10	0.60	0.15	0.20
	3	0.0	0.0	0.05	0.15	0.05	0.05
T_B	1				100.0	100.0	100.0
	2				100.0	100.0	100.0
	3				100.0	100.0	100.0
T_{B1}	1				1.70	0.0	0.0
	2				0.65	0.0	0.0
	3				0.45	0.0	0.0
T_{B2}	1				16.0	0.0	12.0
	2				10.0	0.0	8.6
	3				9.95	0.0	12.0
T_{ADF}	1				100.0		100.0
	2				100.0		100.0
	3				100.0		100.0
T_{ADF1}	1				1.85		0.0
	2				0.40		0.20
	3				0.20		0.30
T_{ADF2}	1				14.0		1.0
	2				7.30†		0.40
	3				4.60‡		0.70

Note. Type I error rates are computed across 2000 replications at the 5% level of significance. D=distributional form: distribution 1=multivariate normal; distribution 2=elliptically symmetric nonnormal; distribution 3=asymmetric nonnormal. † indicates rejection rate within the robustness interval; ‡ indicates rejection rate within the 95% CI.

Table 4. Type I Error Rates for Properly Specified Model B1 (196 df)

n-cond:		1:1	2:1	5:1	10:1	df+1	p*
n:		35	70	175	350	197	231
	D						
T_{ML}	1	92.0	38.0	12.0	8.05	12.0	9.35
	2	99.0	94.0	88.0	88.0	88.0	89.0
	3	100.0	100.0	100.0	100.0	100.0	100.0
$T_{ML\text{-}k}$	1	0.15	0.55	2.45	3.75†	3.20†	2.90†
	2	8.85	45.0	74.0	82.0	76.0	78.0
	3	91.0	100.0	100.0	100.0	100.0	100.0
T_{SB1}	1	97.0	50.0	15.0	9.50	14.0	11.0
	2	96.0	46.0	15.0	9.80	14.0	12.0
	3	94.0	43.0	15.0	9.85	14.0	13.0
$T_{SB1\text{-}k}$	1	0.85	1.60	3.30†	4.50‡	3.90†	3.70†
	2	0.80	1.75	3.50†	5.05‡	4.05‡	5.0‡
	3	1.45	2.25	4.60‡	5.55‡	5.60‡	5.50‡
T_{SB2}	1	29.0	4.25‡	2.55†	3.10†	3.05†	2.15
	2	7.95	0.45	0.35	0.75	0.20	0.30
	3	2.20	0.05	0.05	0.0	0.05	0.05
$T_{SB2\text{-}k}$	1	0.0	0.0	0.30	1.15	0.35	0.35
	2	0.0	0.0	0.0	0.40	0.0	0.05
	3	0.0	0.0	0.0	0.0	0.0	0.0
T_B	1				100.0	100.0	100.0
	2				100.0	100.0	100.0
	3				100.0	100.0	100.0
T_{B1}	1				2.70†	0.0	0.0
	2				13.0	0.0	0.0
	3				100.0	0.0	0.0
T_{B2}	1				18.0	0.0	15.0
	2				45.0	0.0	34.0
	3				100.0	0.0	100.0
T_{ADF}	1				100.0		100.0
	2				100.0		100.0
	3				100.0		100.0
T_{ADF1}	1				3.50†		0.0
	2				15.0		0.0
	3				100.0		0.10
T_{ADF2}	1				19.0		9.55
	2				46.0		18.0
	3				100.0		100.0

Note. Type I error rates are computed across 2000 replications at the 5% level of significance. D=distributional form; distribution 1=multivariate normal; distribution 2=elliptically symmetric nonnormal; distribution 3=asymmetric nonnormal. † indicates rejection rate within the robustness interval; ‡ indicates rejection rate within the 95% CI.

Table 5 Type I Error Rates for Properly Specified Model B2 (102 df)

N:cond:		1:1	2:1	5:1	10:1	df+1	p*
N		129	258	645	1290	103	231
	D						
T_{ML}	1	12.0	9.25	6.30†	6.20†	14.0	7.75
	2	24.0	17.0	17.0	15.0	22.0	19.0
	3	86.0	92.0	96.0	98.0	84.0	91.0
$T_{ML\,A}$	1	2.85†	4.15‡	4.70‡	5.35‡	1.60	3.35†
	2	7.0†	9.55	14.0	14.0	5.40‡	10.0
	3	72.0	87.0	95.0	98.0	66.0	86.0
T_{SB1}	1	15.0	10.0	6.75†	6.30†	18.0	9.15
	2	19.0	10.0	6.95†	5.55‡	20.0	11.0
	3	31.0	17.0	9.30	7.30†	41.0	19.0
$T_{SB1\,A}$	1	3.65†	4.75‡	4.80‡	5.60‡	2.35	3.85†
	2	5.10‡	4.20‡	5.10‡	4.40‡	3.55†	5.35‡
	3	9.20	8.70	7.30†	5.95‡	8.40	9.10
T_{SB2}	1	5.30‡	5.35‡	4.80‡	5.60‡	4.05‡	4.35‡
	2	2.40	2.05	2.90†	2.80†	2.20	2.0
	3	0.40	0.30	0.55	0.85	0.65	0.30
$T_{SB2\,A}$	1	0.65	2.05	3.75†	4.70‡	0.25	1.60
	2	0.10	1.10	1.75	2.50†	0.20	0.65
	3	0.05	0.05	0.25	0.70	0.05	0.10
T_B	1	100.0	95.0	40.0	18.0	100.0	98.0
	2	100.0	96.0	39.0	15.0	100.0	99.0
	3	100.0	98.0	39.0	15.0	100.0	99.0
T_{B1}	1	0.0	1.75	4.35‡	4.55‡	0.0	0.90
	2	0.0	0.80	2.60†	3.35†	0.0	0.45
	3	0.0	0.30	2.0	2.0	0.0	0.20
T_{B2}	1	6.95†	6.35†	5.80‡	5.45‡	0.0	6.10†
	2	4.05‡	4.15‡	4.05‡	3.65†	0.0	4.10‡
	3	6.45†	2.85†	3.25†	3.05†	0.0	2.60†
T_{AD9}	1		94.0	39.0	18.0		96.0
	2		94.0	37.0	15.0		93.0
	3		96.0	34.0	13.0		90.0
T_{ADF1}	1		0.90	4.30‡	4.50‡		0.50
	2		0.10	2.45	3.35†		0.20
	3		0.0	1.15	1.70		6.45†
T_{ADF2}	1		3.35†	5.75‡	5.50‡		2.10
	2		0.75	3.60†	3.50†		1.65
	3		0.25	2.15	2.30		13.0

Note. Type I error rates are computed across 2000 replications at the 5% level of significance. D=distributional form; distribution 1=multivariate normal; distribution 2=elliptically symmetric nonnormal; distribution 3=asymmetric nonnormal. † indicates rejection rate within the robustness interval; ‡ indicates rejection rate within the 95% CI.

Table 6. Empirical Power Estimates for Misspecified Model A1 (87 df)

		n					
	D	35	70	175	350	86	120
T_{ML}	1				100.0†		
	2						
	3						
$T_{ML\,k}$	1	81.0	100.0	100.0†	100.0†	100.0†	100.0†
	2	84.0	100.0†		100.0†	100.0†	100.0†
	3						
T_{SB1}	1				100.0†		
	2				100.0†		
	3						
$T_{SB1\,k}$	1	87.0†	100.0†	100.0†	100.0†	100.0†	100.0†
	2	88.0†	100.0†	100.0†	100.0†	100.0†	100.0†
	3	89.0†	100.0†	100.0†	100.0†	100.0†	100.0†
T_{SB2}	1		100.0†	100.0†	100.0†	100.0†	100.0†
	2	77.0†	100.0	100.0	100.0	100.0	100.0
	3	50.0†	80.0	100.0	100.0	90.0	97.0
$T_{SB2\,k}$	1	23.0	100.0	100.0	100.0†	100.0	100.0
	2	14.0	95.0	100.0	100.0	100.0	100.0
	3	4.60	56.0	99.0	100.0	78.0	94.0
T_B	1						
	2						
	3						
T_{B1}	1			98.0	100.0†		0.0
	2			74.0	100.0†		0.0
	3			53.0	100.0		0.20
T_{B2}	1						
	2			98.0†	100.0†		54.0†
	3			91.0†	100.0†		53.0†
T_{ADF}	1						
	2						
	3						
T_{ADF1}	1			93.0	100.0†		
	2			30.0	100.0†		
	3			8.90	100.0		
T_{ADF2}	1			100.0†			
	2			79.0	100.0†		
	3			49.0	100.0		

Note. All power estimates are computed across 1000 replications. D=distributional form; distribution 1=multivariate normal; distribution 2=elliptically symmetric nonnormal; distribution 3=asymmetric nonnormal. † indicates corresponding Type I error rate within the robustness interval.

Table 7. Empirical Power Estimates for Misspecified Model A2 (110 df)

		n					
	D	18	36	90	180	103	120
T_{ML}	1						
	2						
	3						
T_{ML-k}	1	20.0	92.0	100.0†	100.0†	100.0†	100.0†
	2	47.0					
	3						
T_{SB1}	1						
	2						
	3						
T_{SB1-k}	1	50.0	96.0†	100.0†	100.0†	100.0†	100.0†
	2	29.0	79.0	100.0†	100.0†	100.0†	100.0†
	3	22.0	49.0†	93.0†			
T_{SB2}	1		92.0†	100.0†	100.0†	100.0†	100.0†
	2		39.0	92.0	99.0	96.0	98.0
	3		5.50	36.0	79.0	46.0	53.0
T_{SB2-k}	1	0.10	32.0	100.0	100.0	100.0	100.0
	2	0.0	2.30	89.0	99.0	93.0	96.0
	3	0.10	0.10	21.0	76.0	34.0	44.0
T_B	1						
	2						
	3						
T_{B1}	1				96.0		0.0
	2				63.0		0.0
	3				14.0		0.0
T_{B2}	1						
	2						
	3						
T_{ADF}	1						
	2						
	3						
T_{ADF1}	1				97.0		
	2				65.0		
	3				14.0		
T_{ADF2}	1						
	2				98.0†		
	3				80.0†		

Note. All power estimates are computed across 1000 replications. D=distributional form; distribution 1=multivariate normal; distribution 2=elliptically symmetric nonnormal; distribution 3=asymmetric nonnormal. † indicates corresponding Type I error rate within the robustness interval.

Table 8. Empirical Power Estimates for Misspecified Model B1 (217 df)

		n					
	D	35	70	175	350	197	231
T_{ML}	1						
	2						
	3						
$T_{ML\text{-}k}$	1	80.0	100.0	100.0	100.0†	100.0†	100.0†
	2						
	3						
T_{SB1}	1						
	2						
	3						
$T_{SB1\text{-}k}$	1	89.0	100.0	100.0†	100.0†	100.0†	100.0†
	2	62.0	99.0	100.0†	100.0†	100.0†	100.0†
	3	25.0	69.0	99.0†	100.0†	99.0†	100.0†
T_{SB2}	1		100.0†	100.0†	100.0†	100.0†	100.0
	2		77.0	99.0	100.0	100.0	100.0
	3	3.50	6.60	60.0	94.0	70.0	80.0
$T_{SB2\text{-}k}$	1	2.30	100.0	100.0	100.0	100.0	100.0
	2	0.10	39.0	98.0	100.0	100.0	100.0
	3	0.0	1.0	50.0	94.0	63.0	76.0
T_B	1						
	2						
	3						
T_{B1}	1				100.0†		0.0
	2						0.0
	3						0.0
T_{B2}	1						
	2						
	3						
T_{ADF}	1						
	2						
	3						
T_{ADF1}	1				100.0†		
	2						
	3						
T_{ADF2}	1						
	2						
	3						

Note. All power estimates are computed across 1000 replications. D=distributional form; distribution 1=multivariate normal; distribution 2=elliptically symmetric nonnormal; distribution 3=asymmetric nonnormal. † indicates corresponding Type I error rate within the robustness interval.

Table 9. Empirical Power Estimates for Misspecified Model B2 (117 df)

		n					
	D	129	258	645	1290	103	231
T_{ML}	1			100.0†	100.0†		
	2						
	3						
$T_{ML\text{-}k}$	1	100.0†	100.0†	100.0†	100.0†	100.0	100.0†
	2	100.0†				100.0†	
	3						
T_{SB1}	1			100.0†	100.0†		
	2			100.0†	100.0†		
	3				100.0†		
$T_{SB1\text{-}k}$	1	100.0†	100.0†	100.0†	100.0†	100.0	100.0†
	2	100.0†	100.0†	100.0†	100.0†	100.0†	100.0†
	3			100.0†	100.0†		
T_{SB2}	1	100.0†	100.0†	100.0†	100.0†	100.0†	100.0†
	2	100.0	100.0	100.0†	100.0†	100.0	100.0
	3	94.0	99.0	100.0	100.0	90.0	99.0
$T_{SB2\text{-}k}$	1	100.0	100.0	100.0†	100.0†	100.0	100.0
	2	100.0	100.0	100.0	100.0†	99.0	100.0
	3	88.0	99.0	100.0	100.0	73.0	99.0
T_B	1						
	2						
	3						
T_{B1}	1	0.0	100.0	100.0†	100.0†		100.0
	2	0.0	97.0	100.0†	100.0†		87.0
	3	0.0	94.0	100.0	100.0		80.0
T_{B2}	1	26.0†	100.0†	100.0†	100.0†		100.0†
	2	16.0†	100.0†	100.0†	100.0†		99.0†
	3	26.0†	99.0†	100.0†	100.0†		97.0†
T_{ADF}	1						
	2						
	3						
T_{ADF1}	1		100.0	100.0†	100.0†		
	2		93.0	100.0	100.0†		
	3		64.0	100.0	100.0		
T_{ADF2}	1		100.0†	100.0†	100.0†		
	2		100.0	100.0†	100.0†		
	3		92.0	100.0	100.0		

Note. All power estimates are computed across 1000 replications. D=distributional form; distribution 1=multivariate normal; distribution 2=elliptically symmetric nonnormal; distribution 3=asymmetric nonnormal. † indicates corresponding Type I error rate within the robustness interval.

Figure 1. Population Model A and Correctly Specified Models A1 and A2

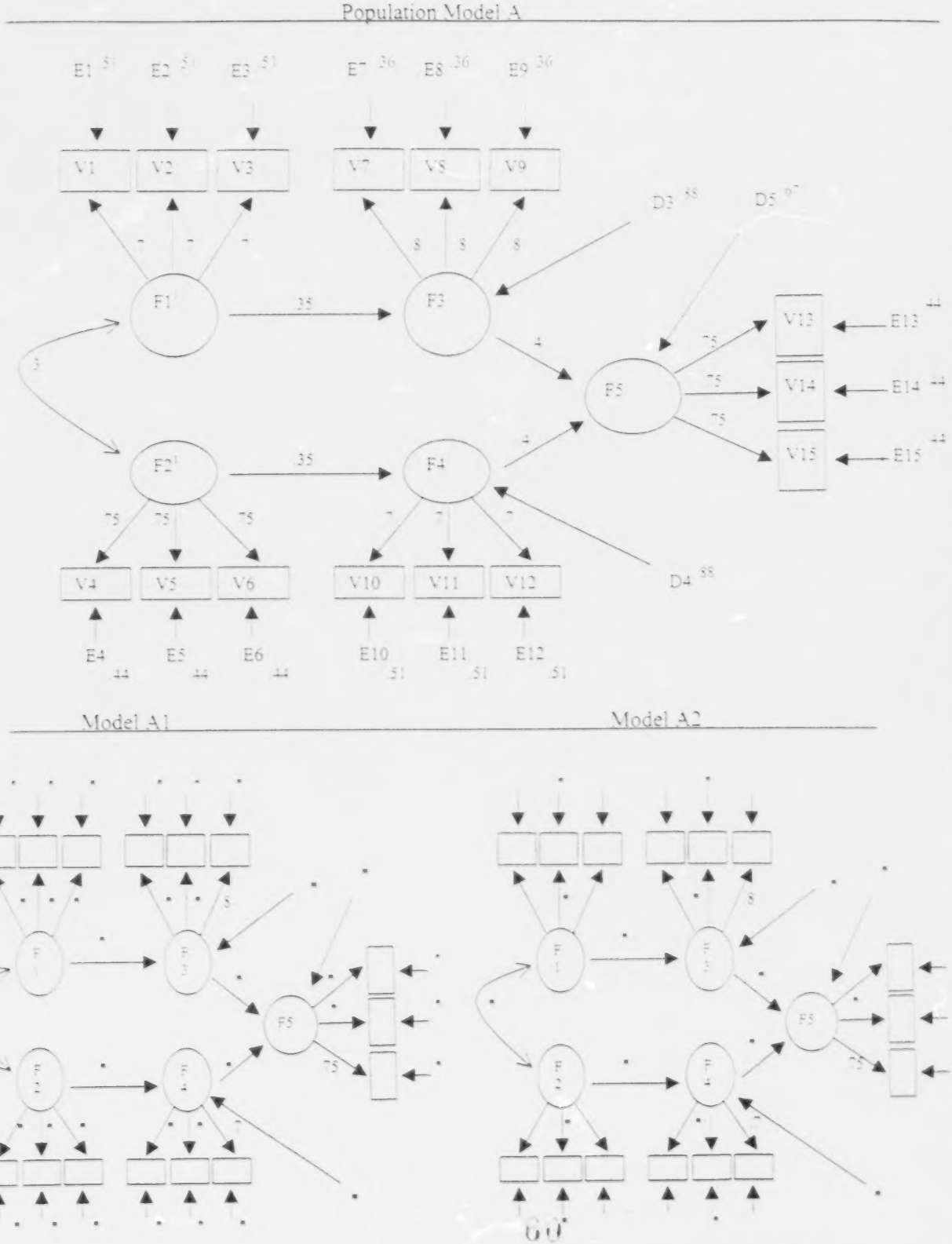


Figure 2. Population Model B and Correctly Specified Models B1 and B2

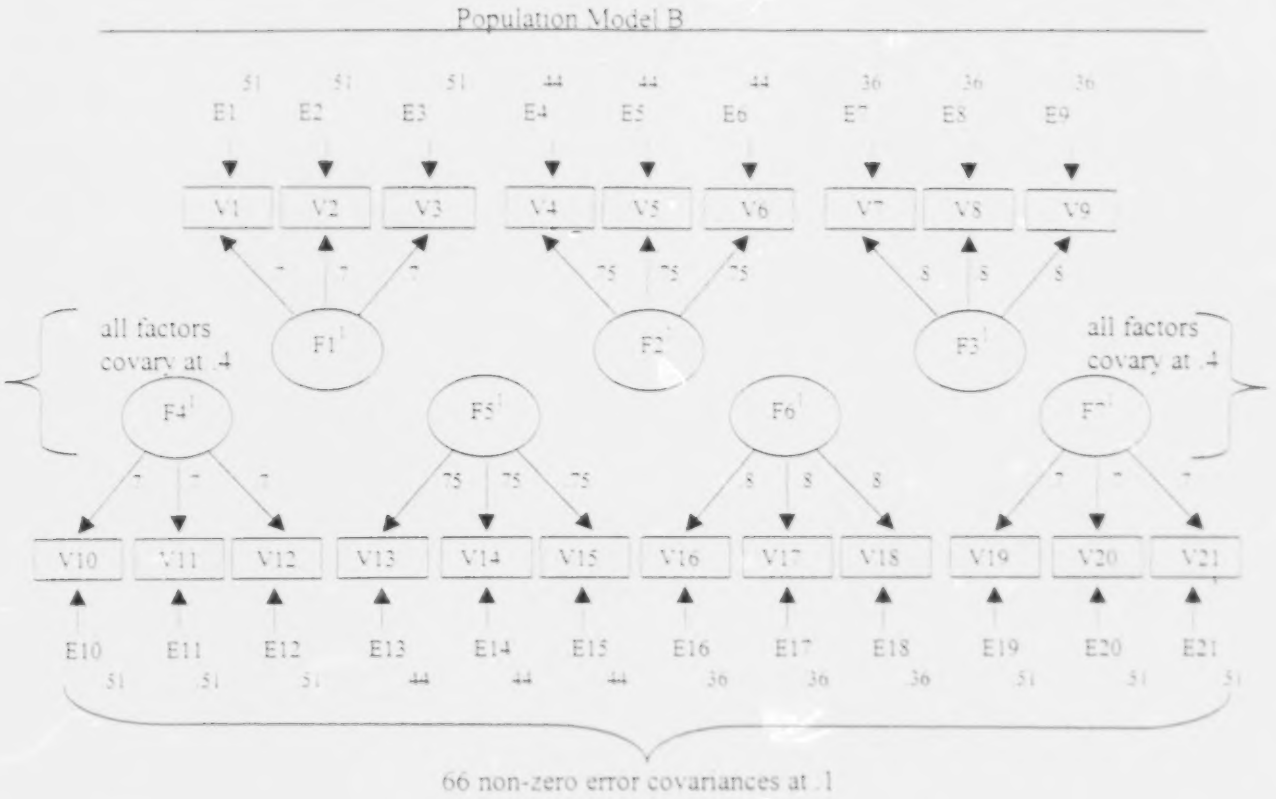

Population Model B
all factors covary at .4
all factors covary at .4
66 non-zero error covariances at .1
Model B1
Model B2
66 error covariances fixed at .1
66 * for error covariances

Figure 3. Misspecified Models A1 and A2 for Power Analysis

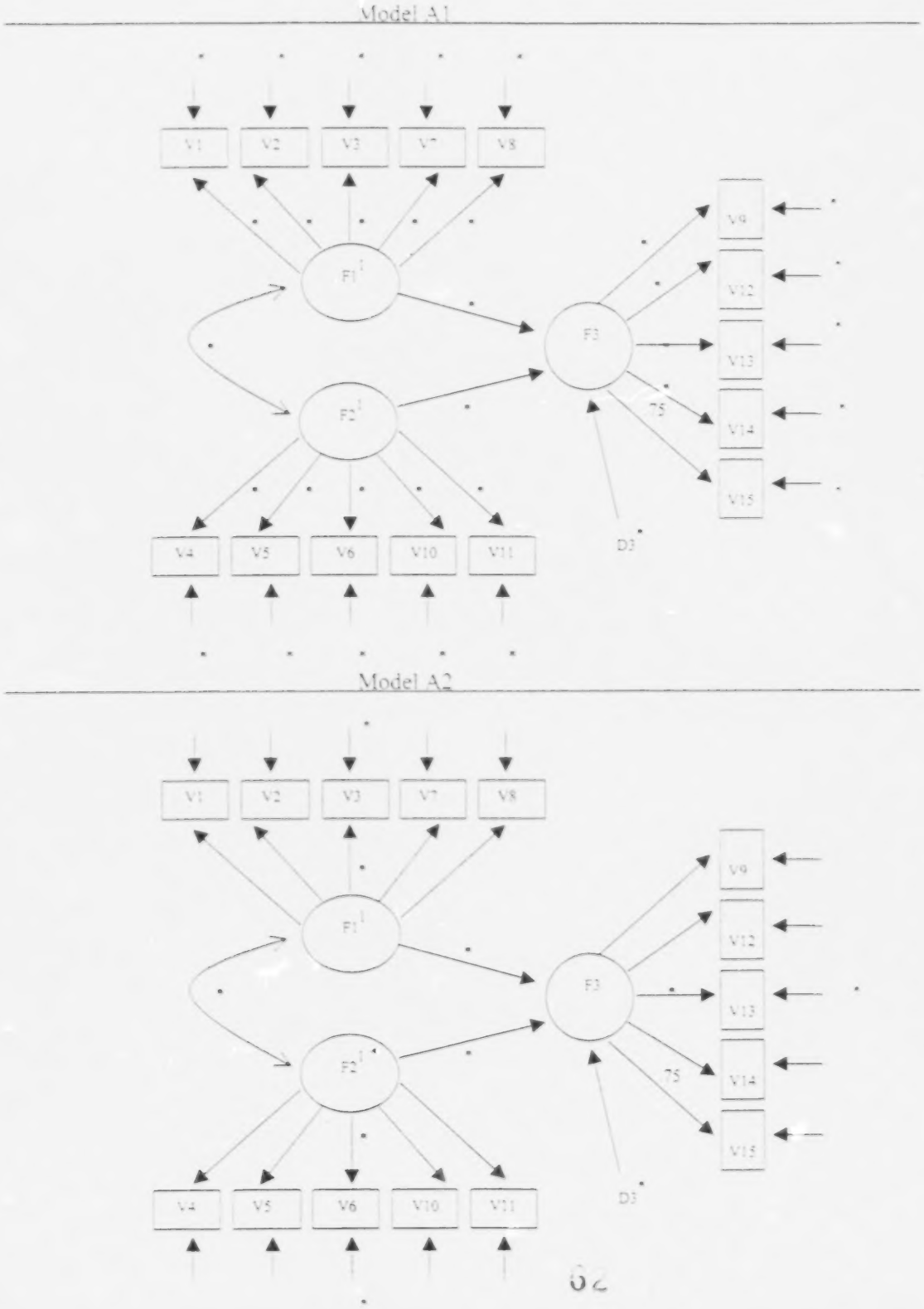
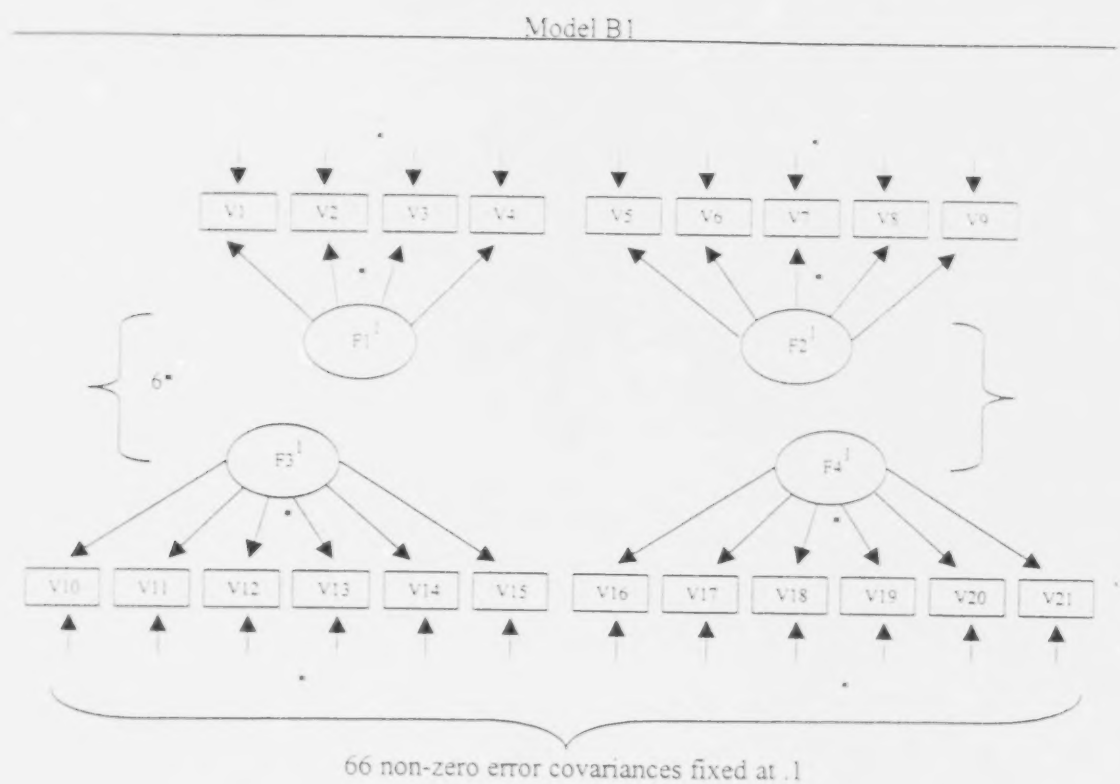
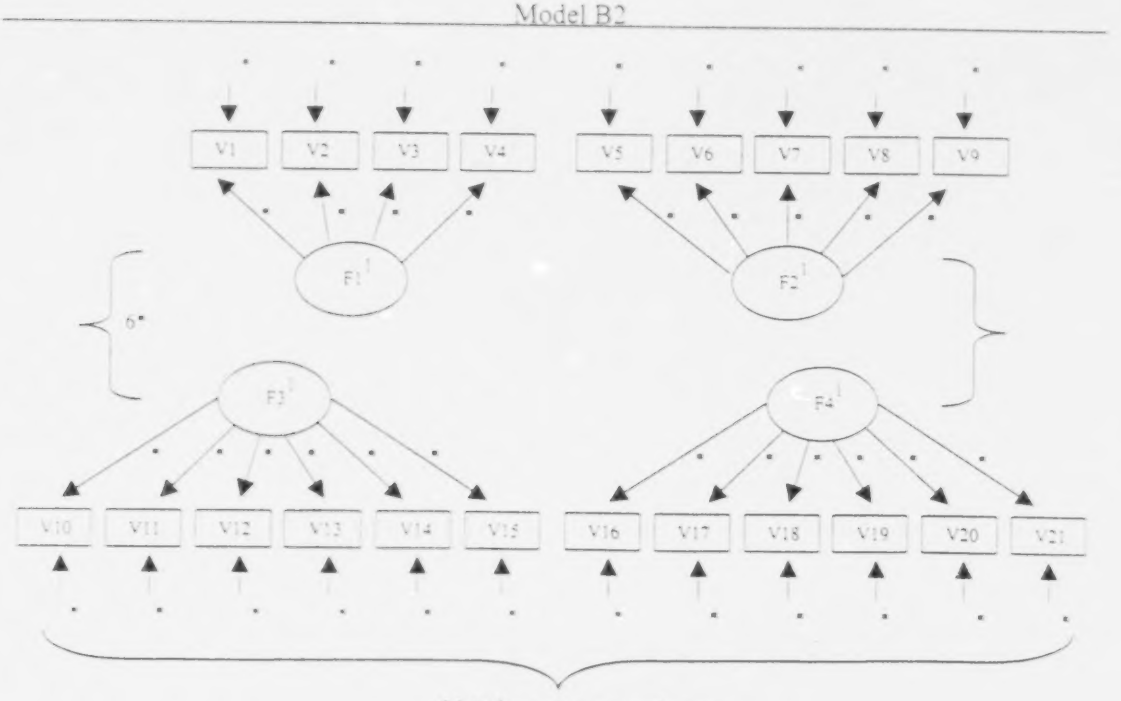


Figure 4. Misspecified Models B1 and B2 for Power Analysis





BEST COPY AVAILABLE